Examination of the Effectiveness of Mental Health First Aid Among the Public:

A Meta-Analysis

A Plan B Research Project
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

AMY K. MASLOWSKI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Rick LaCaille, Ph.D., Advisor
Lara LaCaille, Ph.D., Committee member
Catherine Reich, Ph.D., Committee member
Jill Klingner, Ph.D., Committee member

May 2018

Abstract

Mental health conditions are prevalent and can significantly impact a person's functioning; however, it can be challenging for an untrained individual to know when or how to provide assistance to a person experiencing a mental health concern. Mental Health First Aid (MHFA) was developed to provide people with training and guidelines to improve their mental health literacy, decrease stigma, and increase their confidence and willingness to offer aid to individuals experiencing a mental health disorder or crisis. Given recent initiatives to expand the implementation of MHFA programs and increased efforts to evaluate it with more rigorous designs, a comprehensive and systematic review of the evidence base is warranted. This study investigated and quantified, via meta-analytic methodology, MHFA outcomes (i.e., knowledge, attitude, behavior) for the training participant as well as both a quantitative and narrative review for the individual experiencing a mental health disorder or crisis. We conducted a comprehensive search for eligible studies that utilized either a control or comparison group. This resulted in 15 studies for inclusion in the meta-analysis. Small-to-moderate effect sizes were found for the primary outcomes for the trainees; however, preliminary evidence suggests effects for the recipients were more difficult to observe. Study quality had a significant moderating effect. Overall, MHFA appears to be an effective intervention for increasing knowledge about mental health, decreasing stigma and social distance, and increasing trainees' confidence in approaching and providing aid to an individual experiencing a mental health disorder or crisis. Considerably greater attention and effort in demonstrating effects on recipients is needed with future empirical investigations.

Table of Contents

List of Tables

List of Figures

Literature Review

Mental health disorders are prevalent worldwide, with one in four adults meeting criteria for a disorder in a given year and about 30% meeting criteria for a condition in their lifetime (Steel et al., 2014). Perhaps not surprisingly, mental health disorders are also associated with considerable hardship and disability. It is estimated that mental illness costs the United States $193.2 billion in lost earnings every year (Insel, 2008). In fact, data from 2010 demonstrated that mental health and substance use disorders were the leading disorders contributing to global years lived with disability, accounting for 183.9 million disability-adjusted life years at 18% (Whiteford et al., 2013). Notably, mental illnesses were more common than cardiovascular and circulatory diseases, neoplasms, and musculoskeletal disorders, which accounted for 16% of the disability-adjusted life years.

Although mental illnesses are common and burdensome, it is often a difficult topic for people to discuss, particularly for those who have a mental illness, for fear of the associated stigma or social rejection (Feldman & Crandall, 2007). Furthermore, adults with mental illnesses are often perceived by the general public as dangerous and burdensome, especially those with drug and alcohol use disorders, depression, and schizophrenia (Parcesepe & Cabassa, 2013). Similarly, children with mental health disorders bear the burden of stigma and are viewed as lazier than children with a physical health diagnosis (e.g., asthma; Parcesepe & Cabassa, 2013). Although there are treatment options for people experiencing mental health concerns, research has estimated that as few as 11% receive treatment over a year's time (Wang, Aguilar-Gaxiola et al., 2007). Many factors influence people's ability or willingness to seek treatment, such as their fear of negative reactions or

evaluations from others. In some cases, people may not be aware that they are experiencing clinically significant symptomatology.

Mental Health First Aid

Mental Health First Aid (MHFA) is a manualized program aimed at improving mental health literacy, reducing negative attitudes and stigma associated with mental illness, and increasing confidence and helping behaviors regarding mental health (Kitchener & Jorm, 2002a, 2002b). Although other mental health intervention programs (e.g., gatekeeper trainings) have been developed over the years, MHFA training has been developed with the intention of being widely disseminated and emphasizing an open and accessible curriculum for the non-mental health professional. Training participants (i.e., trainees) have included first responders (e.g., police officers, paramedics, firefighters), teachers or professors, coaches, healthcare personnel (e.g., doctors, nurses, assistants), resident advisors and directors, and family members of people with mental health disorders. MHFA has been formatted into an 8- or 12-hour training and as either adult or youth-focused. In both formats, participants learn an *Action Plan* methodology and accompanying skills, including learning how to assess an individual's risk of suicide or harm, listening non-judgmentally, giving reassurance and information, encouraging an individual to seek the appropriate help, and encouraging an individual to utilize other social supports and self-help strategies (Kitchener & Jorm, 2002a, 2002b). Trainees can expect to learn how to respond in a crisis situation involving suicidal thoughts or behaviors, panic attacks, and psychotic behavior. More common mental illnesses such as depression, anxiety, and psychotic disorders are also described, in addition to substance use disorders, which are commonly comorbid (Kitchener & Jorm, 2002a, 2002b). The Youth Mental Health First Aid (YMHFA) training is similar to

the adult course but is modified to be more age-appropriate and specific to the needs of an adolescent population (Aakre, Lucksted, & Browning-McNee, 2016; Kelly et al., 2011; Gryglewicz, Childs, & Soderstrom, 2018).

Both the youth and adult-focused MHFA training programs entail an interactive curriculum with the main objectives of 1) increasing the trainees' mental health knowledge and 2) helping behaviors and 3) decreasing stigmatizing attitudes toward people with mental health disorders (Kitchener & Jorm, 2002a, 2002b). The knowledge outcome involves mental health literacy, which includes recognizing common mental health disorders, awareness of treatment options and self-help strategies, and first aid skills to be used in a mental health crisis (Kitchener & Jorm, 2002a, 2002b). Proper identification of a mental illness is a crucial first step for trainees to use their MHFA training effectively. Unfortunately, under-recognition of mental illnesses is common though variable across countries (Dahlberg, Waern, & Runeson, 2008; Jorm & Kitchener, 2011; Jorm et al., 2005; Pescosolido et al., 2008; Wang, Adair et al., 2007). Furthermore, mental health literacy influences people's attitudes toward mental health (Kelly, Jorm, & Wright, 2007).

The mental health literacy and knowledge focus within the MHFA curriculum is thought to facilitate the reduction of trainees' stigmatizing attitudes about mental health that can interfere with the prospective benefits of someone providing social support (Cheng, Wang, McDermott, Kridel, & Rislin, 2018; Jung, Sternberg, & Davis, 2017; Kelly, Jorm, & Wright, 2007; Swarbrick & Brown, 2015). The attitude outcome involves reducing trainees' personal and perceived stigma (Kitchener & Jorm, 2002a, 2002b). Personal stigma has been conceptualized as one's own beliefs about mental illnesses, whereas perceived stigma assesses trainees' beliefs about what other people believe about mental illnesses (Jorm,

Kitchener, Sawyer, Scales, & Cvetkovski, 2010b). The literature has been mixed about whether changes in perceived stigma increases or decreases as a result of increased education, therefore warranting further review (National Academy of Sciences, Engineering, and Medicine, 2016). However, perceived stigma is closely tied with mental health literacy and has been measured in many MHFA evaluations (e.g., Burns et al., 2017; Jensen, Morthorst, Vendsborg, Hjorthøj, & Nordentoft, 2016; Jorm, Kitchener, Fischer, & Cvetkovski, 2010a; Jorm et al., 2010b; Svensson & Hansson, 2014). It is presumed that increases in knowledge about mental illnesses will result in changes in one's attitudes. Furthermore, this outcome involves social distance, which is often measured to assess trainees' avoidance or tolerance with people who have mental health disorders (i.e., recipients; Griffiths, Christensen, Jorm, Evans, & Groves, 2004; Link, Phelan, Bresnahan, Stueve, & Pescosolido, 1999). Once trainees learn about mental health literacy and are more cognizant about their stigmatizing attitudes, the final step is for them to provide aid to people in need.

Providing actual help to a person in need is one of the main purposes of MHFA training. Typically, providing aid has been operationalized by trainees rating their confidence with helping someone, the number of times they have had contact with an individual with a mental illness (i.e., recipients), if they offered help (i.e., taking time to listen nonjudgmentally; asking about suicidal thoughts; giving information about treatment options; suggesting self-help strategies), and if they advised or referred an individual to professional services. Often researchers have relied upon self-report of helping behaviors via post-training surveys (Jensen et al., 2016; Jorm, Kitchener, O'Kearney, & Dear, 2004; Kitchener & Jorm, 2002a, 2002b, 2004; Svensson & Hansson, 2014).

Finally, although not a primary outcome of MHFA, the trainees' own psychological distress both before and after the intervention has been measured in several studies to determine the impact of the program on their well-being (Davies, Beever, & Glazebrook, 2018; Jorm et al., 2010a, 2010b; Kitchener & Jorm, 2004; Lipson, Speer, Brunwasser, Hahn, & Eisenberg, 2014).

MHFA Implementation in the U.S.

The MHFA curriculum was originally developed in Australia (Kitchener & Jorm, 2002a) and has been implemented and evaluated in other parts of the world such as England, Finland, Hong Kong, and Canada, but the specific use of MHFA is a newer phenomenon within the United States. In 2012, the Substance Abuse and Mental Health Services Administration (SAMHSA) added MHFA to its National Registry of Evidence-Based programs; however, the three trials that were reviewed, were all authored by the MHFA founders (SAMHSA, 2012). Notably, following recent mass shootings and crimes (for example in Parkland, FL, 2018; Newtown, CT, 2012; Aurora, CO, 2012) that may have, in part, stemmed from underlying mental health conditions, the U.S. government placed an emphasis on funding MHFA training for widespread implementation (National Council for Behavioral Health, 2014; White House, President Barack Obama, 2013). Furthermore, in 2014, legislation was enacted in more than 23 states across the U.S. with more than $35 million allocated towards MHFA interventions (National Council for Behavioral Health, 2014). Recently, one U.S. celebrity started a campaign to train people along her tour schedule – totaling an additional 150,000 MHFA trainees (Rosenberg & Germanotta, 2018). Clearly, the presence and attention to training people in MHFA within the U.S. is increasing.

Given the emphasis on disseminating MHFA, it is important to systematically review and summarize the evidence for this training program in the U.S. (as well as in other countries). MHFA – USA has some modifications that may meaningfully differentiate it from the Australian version of MHFA. MHFA – USA is only offered as an 8-hour course, but it still incorporates the same coverage of skills taught in the 12-hour course. Furthermore, cultural differences between the U.S. and Australia, such as beliefs about the etiology of certain disorders, could influence the effectiveness of the training (Aakre, Lucksted, & Browning-McNee, 2016; Morgan, Reavley, Jorm, & Beatson, 2017). Therefore, reviewing the MHFA evidence for both MHFA – USA as well as MHFA in other countries will help identify if any programmatic differences in effectiveness are present.

Previous Reviews of MHFA Outcomes

Given the enthusiasm and resources allocated for MHFA programming, it is not surprising that greater efforts to systematically review the evidence base for these trainings has occurred. An early narrative review by Kitchener and Jorm (2006) of three of their MHFA evaluation studies (all randomized control trials) suggested that trainees experienced improved knowledge about mental illnesses, greater confidence in providing help, increased self-reported helping behavior, and decreased stigmatizing attitudes towards mental illness over wait-list controls. Long-term effects of the training were reported to be maintained for about five to six months following the trainings.

More recently, Booth and colleagues (2017) published a systematic review of mental health training programs, which included three MHFA studies, as well as other interventions with a similar focus. The review included 19 total studies, of which 15 were randomized control trials (RCTs) or non-randomized control trials (nRCTs); the primary trainee

population focus was police officers and others who often come in contact with people experiencing a mental health crisis. Primary outcomes for the review included examining changes in the trainees' knowledge and helping behavior, as well as to identify methods that have been utilized to assess recipient effects, though only three of the included studies reported recipient data. Booth and colleagues (2017) briefly summarized the findings of the three MHFA studies but did not directly compare results among the different types of interventions; it may have been helpful to see a more direct comparison of the effectiveness of several active treatments. Furthermore, this review may have overlooked several other eligible MHFA studies for inclusion in the review (e.g., Kitchener & Jorm, 2004; Massey, Brooks, & Burrow, 2014; Moffitt, Bostock, & Cave, 2014). Nevertheless, the researchers concluded that, overall, the trainings yielded short-term positive benefits. However, there was little to no detected effect for the recipients with whom the trainees came into contact with following the training.

To date, only one MHFA-exclusive meta-analysis has been published. Hadlaczky, and colleagues (2014) conducted a meta-analysis of 15 MHFA studies (four RCTs, two nRCTs, and nine pre/post evaluations without control groups) evaluating knowledge, attitude, and behavioral outcomes for trainees. The review suggested that MHFA had small-to-moderate sized effects for each of the targeted outcomes (knowledge Glass's $\Delta = 0.56$; attitude Glass's $\Delta = 0.28$; behavior Glass's $\Delta = 0.25$). Importantly, considerably smaller effects were found for the six controlled studies in comparison to the nine uncontrolled studies (knowledge Glass's $\Delta = 0.38$ vs. 0.68; attitude Glass's $\Delta = 0.23$ vs. 0.33; behavior Glass's $\Delta = 0.24$ vs. 0.33); however, the moderating differences of study design were not found to be significant. Also notable, the behavior outcome was based upon self-reported

contact and the number of times the trainees had utilized their MHFA skills with another person, which required the trainees to encounter a person with a mental health disorder or in a mental health crisis. This may have been problematic because that cannot be assessed unless the trainees actively seek out people in need (Hadlaczky, Hökby, Mkrtchian, Carli, & Wasserman, 2014).

Although an important initial step in evaluating MHFA, this meta-analysis had several limitations, including the utilization of a large number of single group design studies (i.e., no comparison group), no studies conducted in the United States, and only analyzing the effects on trainees, rather than for both the trainees and recipients. Moreover, Hadlaczky and colleagues (2014) only calculated effect sizes based on the differences in scores between the baseline and latest follow-up, not any of the interim time points. This appears problematic because it limits the identification of changes that have occurred and if MHFA has a more enduring effect on the trainees. Importantly, the meta-analysis also lacked a well-validated method of reviewing study quality, as quality was operationalized as the publishing journal's impact factor. This study quality proxy is problematic, as several factors may influence a journal's impact factor independent of a study's rigor and quality. Lastly, since 2014, MHFA has increasingly been disseminated and researched with more controlled and rigorously designed studies; thus, an update to the literature seems timely.

Recipient Outcomes with MHFA

Approaching at-risk individuals (recipients; those with a mental health condition or experiencing a mental health crisis) with confidence and connecting them with treatment options is an important point of intervention. Thus, the effects of MHFA on the recipients are of the utmost importance. Unfortunately, considerably less research has been conducted

involving recipients. Though limited research has assessed recipient effects, at least two RCTs (Jorm et al., 2010b; Lipson et al., 2014) have directly measured changes, and at least one single group pre/post study (Pierce, Liaw, Dobell, & Anderson, 2010) has indirectly measured recipient effects. To date, research on recipient effects has emphasized whether the individual in crisis talked with a MHFA trained person, whether the trained person listened to their problem, helped them calm down, talked to them about suicidal thoughts, or provided/recommended a referral to professional help (Jorm et al., 2010b; Lipson et al., 2014; Pierce et al., 2010; Wong, Collins, & Cerully, 2015). Recipient effects have typically been collected through follow-up surveys given to people who potentially came into contact with those participating in one of the trainings (e.g., students in a school with trained teachers, athletes on a team with trained coaches). Notably, the nature and type of support provided by the trained individuals may influence whether the recipients report the impact of the intervention (Wong, Collins, & Cerully, 2015) and warrants greater empirical investigation.

Previous Research Limitations

Two major factors have sparked concerns regarding the effectiveness of MHFA. First, MHFA research has less consistently examined recipient outcomes, even though people with mental illnesses or those in crisis are the target population for this intervention. Tracking effects on the recipients is often challenging because it requires long-term follow-up with people who may have been given aid by a trainee (Wong, Collins, & Cerully, 2015). In their 2006 evaluation, Kitchener and Jorm affirmed that recipient effects had not yet been measured and were the necessary next step. Researchers have begun to examine the effects on recipients within certain settings, such as schools, where outcomes may be more readily

directly assessed. A second, though perhaps less concerning limitation, is that the primary body of MHFA research has been conducted by the program's founders (Graham, 2013). Kitchener and Jorm have consistently found positive results with the program's effectiveness and have published over 35 studies since 2001. More recently, however, other researchers unaffiliated with the founders have also conducted equally as controlled and rigorously designed research, and they appear to also report positive effects for the trainees.

Objectives

Given the resources being allocated for MHFA intervention dissemination, a strong evidence base for the proposed trainee and recipient outcomes is warranted. The current meta-analytic study evaluated and systematically reviewed the most current MHFA evidence available through an exclusive utilization of randomized and non-randomized control trials. Specifically, this study focused on the trainees' mental health literacy (e.g., knowledge of effective treatments for common conditions and ability to correctly identify a condition), attitudes (e.g., reduced stigma towards people with mental health conditions), and helping-related behavior (e.g., confidence with approaching and giving aid to someone in need), as well as the impact of the program (e.g., increased help seeking, reduced psychological distress) for the people who came into contact with trainees (i.e., recipients). In addition, the analysis examined changes in the trainees' psychological distress, but due to the limited number of studies reporting this, we considered it a secondary outcome. We also examined the potential moderating effects of study quality or risk of bias, study design, investigator allegiance, country of research, and trainee gender. Moreover, with the growing body of literature on MHFA, this meta-analysis helps to consolidate the findings and identify gaps remaining in the literature.

Method

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009) were used for the current meta-analytic review. The 27-item checklist (see Appendix A) provides an overview of our process and methodology.

Search Strategy and Study Selection

We completed a comprehensive literature search to gather all relevant empirical studies meeting inclusion criteria. This included 1) searching in online databases for both published studies and dissertations or theses, 2) searching online databases within 18 journals (e.g., *BMC Psychiatry, International Journal of Mental Health*; see Appendix B for full list of journals) that commonly publish research in this area, 3) searching reference sections of included studies and studies of interest, 4) searching the MHFA websites (https://mhfa.com.au; https://www.mentalhealthfirstaid.org), and 5) contacting five researchers who had published protocols for MHFA intervention studies. Search terms used were: *mental health first aid*, *MHFA*, *mental health training*, *mental health gatekeeper training*, *mental health gatekeeper*, *mental health education*, and *mental health literacy* from the following databases: PsycINFO, PsycARTICLES, Psychology Database, Consumer Health Complete, PubMed, and Scopus (see Appendix B for search syntax). We included studies that were published any time before March 2018, with the oldest study published in 2010.

Included studies (a) were randomized or non-randomized control trials (e.g., quasi-experimental design), (b) based upon the adult or youth MHFA curriculum that involved either an 8- or 12-hour training protocol, (c) published in English, and (d) published in

scientific, peer reviewed journals or were dissertations or theses. Excluded MHFA intervention studies were primarily those that 1) utilized a single group pre-posttest evaluation design, 2) provided only a narrative or qualitative review, or 3) used an incomplete version of the MHFA curriculum.

Two authors independently reviewed all of the titles and abstracts from the initial database search, as well as the full-text articles that were selected for further review. Satisfactory levels of reliability were established between the researchers with an overall rater agreement of 99.7% and kappa agreement of .70. See Table 1 for a breakdown of the included and excluded studies by rater. Disagreements regarding inclusion and coding were resolved through discussion among the study authors.

Risk of Bias

The Cochrane Collaboration's tool for assessing risk of bias was used to assess the quality of the included studies (Higgins et al., 2011; see Appendix C). This tool was developed for use with randomized control trials, but we used it to assess bias for all included studies. The Cochrane tool incorporates bias ratings of high, low, and uncertain across seven domains: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other bias. The blinding of participants and personnel item was excluded from our total quality rating because it was not feasible given the nature of the interventions examined in this meta-analysis. The incomplete outcome data item assessed the studies' use of intent-to-treat analyses to account for attrition. When both intent-to-treat and non-intent-to-treat data were available, we used the non-intent-to-treat data.

Bias was assessed and agreed upon by two of the authors for each included study. We determined the overall threshold levels for high (i.e., two or more individual items with high bias ratings) and low (i.e., one or fewer individual items with a high bias rating) bias studies and provided each study with a total risk of bias score (Higgins & Altman, 2008). If we were unable to determine the bias level of an item for a study, we rated the item as uncertain; these items did not contribute to the overall bias rating for the study.

Outcome Measures

Knowledge, attitude, and behavior were assessed in a variety of ways across studies and sometimes through more than one measure; therefore, where conceptually congruent, we combined outcome measures to determine overall effects for each domain. Most measures were collected prior to (i.e., baseline) and following the MHFA training (i.e., post), though the post-intervention assessment periods varied from immediately afterwards to weeks or months later.

Knowledge. Trainee mental health literacy was often assessed via a true or false response-format test (e.g., Jorm et al., 2010b) and/or by asking individuals to identify common mental health disorders (i.e., depression; schizophrenia) based upon a written vignette. The recipients' knowledge was assessed through disorder identification only. The validity of these measures has not been assessed.

Attitude. Trainee attitudes were most commonly assessed through validated self-report scales (e.g., Social Distance Scale; Link et al., 1999) and/or changes in personal and/or perceived stigma (e.g., Griffiths et al., 2004). In some cases, this outcome was considered in relation to the vignettes described above (e.g., "How likely would you be to have John as a co-worker?" or "How dangerous do you think John is?"). If vignettes were not used, attitude

was assessed as an independent measure (e.g., “How dangerous do you think others believe an individual with a mental health disorder is?”). Both personal and perceived stigma, as well as social distance, were pooled together to calculate our total attitude effect size. Given the mixed literature on changes in perceived stigma, we also calculated a total attitude outcome effect size excluding perceived stigma data, to assess differences that may exist. The recipients were only asked about personal stigma.

Behavior. Trainees’ behavior toward potential aid recipients was most commonly operationalized and assessed via the trainees’ confidence in helping someone experiencing a mental health disorder or crisis and/or, in some cases, self-reported aid to a person in need. Although confidence in approaching a person who may be experiencing a mental health condition may not always equate to actual helping behavior, we determined that confidence was an important initial step towards actual behavior. Recipients were only asked about help they had received and/or treatment they had sought. The validity of these measures has also not been assessed.

Distress. Trainee and recipient psychological distress level was assessed through validated measures, including one of the Kessler Psychological Distress Scales (i.e., K6 or K10; Kessler et al., 2002) the Depression Anxiety and Stress Scale (DASS-21; Lovibond & Lovibond, 1995), or a short form health survey (SF-12; Ware, Kosinsk, & Keller, 1996). We did not include distress in any of our overall calculations of effect size because doing so would have limited our comparability to the previous meta-analysis (Hadlaczky et al., 2014). Furthermore, distress was measured in only a limited number of the studies.

Statistical Analyses

Effect sizes were calculated and analyzed with Comprehensive Meta-Analysis Version 3 (CMA-V3; Borenstein, Hedges, Higgins, & Rothstein, 2009). The type of data reported across studies varied, which required the use of several different methods in calculating effect sizes. However, CMA is able to compute effect sizes for approximately 100 different types of data (e.g., means and standard deviations; odds ratios; *F*-values). When possible, we used means and standard deviations; when insufficient data were reported in the published paper to estimate effect sizes, the study authors were contacted for additional data. In some studies, odds ratios were reported for single items within one outcome. In these cases, we first calculated a total odds ratio for the construct and then converted that to our effect size. Hedges' *g* ($(\overline{x}_{\text{exp post}} - \overline{x}_{\text{exp pre}}) - (\overline{x}_{\text{cont post}} - \overline{x}_{\text{cont pre}})/SD_{pooled}$) rather than Cohen's *d* was used as the effect size measure as *g* provides a better effect size estimate when the sample size is small. Effect size magnitude was interpreted based on the standards of small (.20 to .49), medium (.50 to .79), and large (.80 or greater; Hedges & Olkin, 1985).

The CMA software uses the meta-analytic inverse variance weighting technique to establish a precise estimate of the pooled effect size of the random effects. This technique converts study outcomes into standard deviation units, which are then corrected for bias (Borenstein et al., 2009; Johnson, Mullen, & Salas, 1995). A random effects model was used rather than the fixed effects model because of the anticipated presence of heterogeneity and assumption that the true effect sizes would likely vary from study to study and are normally distributed. Under a fixed effects model, it is assumed that there is one true effect and that any variation in that effect is due to sampling error. In a random effects model, each study is assigned a weight in order to calculate the effect size, with the weight based upon its within-study and between-study error (Borenstein et al., 2009).

It is important to assess homogeneity in a meta-analysis, which determines if the observed between-study dispersion in effect sizes is statistically significantly more than would be expected by chance. To assess for homogeneity within our sample of studies, we calculated a *Q* statistic along with the accompanying *p*-value. A significant *Q* statistic reflects heterogeneity. A non-significant *p*-value could reflect homogeneity or may be due to a lack in power, as the *Q* statistic commonly lacks power in meta-analyses with small sample sizes. Therefore, we also calculated I^2, which is not dependent on sample size. An I^2 value of 25% or less suggests small heterogeneity, around 50% suggests moderate heterogeneity, and 75% or greater suggests large heterogeneity (Borenstein et al., 2009). In order to further evaluate the meaning of the heterogeneity, we conducted additional moderation analyses with outcomes that had large I^2 values. Meta-regression was used for this analysis when moderator variables were continuous rather than categorical format. When the variable was in a categorical format, we utilized subgroup analysis within CMA to analyze the potential moderation. A priori moderators included quality of study (i.e., risk of bias ratings), study design (i.e., RCT vs. nRCT), investigator allegiance (i.e., studies in which the MHFA developers were listed as authors), country of research (i.e., studies published in Australia vs. all other locations), and trainee gender (i.e., percent female).

We also assessed for the potential effects of publication bias. Published studies are likely to report significant effects whereas studies that did not find significant effects are more likely to become discarded (i.e., file drawer problem; Rosenthal, 1979). Thus, these effects may overinflate effect sizes summarized within a meta-analysis. Therefore, we assessed for publication bias through funnel plots, trim-and-fill strategies, Fail-safe *N*, and Egger's regression intercept (Duval & Tweedie, 2000; Rothstein, 2005). Funnel plots provide

a graphical means of assessing bias with the x-axis representing the total effect sizes and the y-axis representing the standard error. One would hope that studies (as represented by open circles) would be evenly dispersed both to the left and right of the mean effect size and that no effect sizes are outside of the plot. Egger's regression intercept estimates asymmetry within the data seen in the funnel plot; a significant *p*-value is indicative of publication bias, though the estimate and confidence intervals are more informative (Borenstein et al., 2009). The trim-and-fill method is an iterative procedure that builds upon the use of funnel plots and provides one with the number of studies that are likely missing from the analysis. This is represented by funnel plots that have both the included studies and others not included (most commonly solid circles to the left of the mean or ones reporting negative effects). Moreover, the trim-and-fill procedure estimates an adjusted unbiased effect size (as indicated by an open diamond). Finally, the Fail-safe *N* test, developed by Rosenthal (1979), indicates the number of additional unpublished studies reporting null findings that would be needed in the analysis to reduce the effect size to the point of nonsignificance. It would be noteworthy if this number was large, meaning that it would take many "in the file drawer" null finding studies to make the total effect size not significant (Rothstein, 2005).

Results

Description of Studies

Figure 1 demonstrates the number of studies excluded and included in the meta-analysis across the stages of review. After excluding duplicate studies, studies missing a title or abstract, and studies from the table of contents search that did not fit our inclusion criteria, we initially screened 6,277 titles and abstracts. We selected 21 studies to be reviewed in full-text after excluding 18 more duplicates and 6,238 studies that did not meet our inclusion

criteria. After full-text review, six more studies were excluded due to not including a control group (n = 2), systematic evaluation of MHFA (n = 1), not using the MHFA curriculum (n = 1), only described the protocol for a MHFA study that is still being conducted (n = 1), and a dissertation that was later published and included (n = 1). This resulted in 15 randomized or non-randomized control trials on MHFA to be included in the trainee analyses (n = 8,184). Two of these studies were also included in preliminary recipient analyses (n = 3,623). See Table 2 for a full description of the included studies.

Quality Assessment

Risk of bias was calculated using the Cochrane Collaboration's risk of bias tool (Higgins et al., 2011). See Figures 2 and 3 for individual item ratings by study and category. We rated eight of the 15 studies as having overall low bias. The seven studies rated as having high bias were deemed adequate to remain in the meta-analysis; however, we did examine bias as a potential moderator. The two most frequently rated items in terms of bias were allocation concealment (high = 5 studies, uncertain = 1) and possible contamination assessed through the other bias category (high = 7 studies, uncertain = 3).

Trainee Outcomes

A statistically significant and medium overall effect size combining the knowledge, attitude, and behavior trainee outcomes was found, Hedges' g = 0.34, CI: [0.23, 0.44], p < .001. See Figure 4 for a forest plot representation. Heterogeneity was small to moderate, Q = 20.32, p = 0.12, I^2 = 31.10%. Significant effects were also found for all three primary trainee outcomes, while trainee distress was marginally significant.

Knowledge. The overall effect size for knowledge from baseline through the latest follow-up (including post-intervention) was medium in magnitude (n = 14; Hedges' g = 0.48,

CI: [0.37, 0.60], $p < .001$). Heterogeneity among the included studies ($n = 14$) was small to moderate ($Q = 21.73$, $p = 0.06$, $I^2 = 40.18\%$). Statistically significant changes were observed from baseline to immediately post-intervention, with moderate heterogeneity ($n = 6$; Hedges' $g = 0.56$, CI: [0.29, 0.83], $p < .001$; $Q = 13.01$, $p = 0.02$, $I^2 = 61.57\%$) and from baseline to latest follow-up (not including immediate post-intervention; $n = 11$; Hedges' $g = 0.50$, CI: [0.38, 0.62], $p < .001$; $Q = 19.01$, $p = 0.04$, $I^2 = 47.41\%$). See Figures 5 - 7 for forest plots of each of these time periods.

Attitude. The overall effect size for attitude from baseline through the latest follow-up (including post-intervention) was small ($n = 13$; Hedges' $g = 0.18$, CI: [0.11, 0.26], $p < .001$). Heterogeneity among the included studies was small ($Q = 12.86$, $p = 0.38$, $I^2 = 6.70\%$). Significant changes were also observed from baseline to immediately post-intervention, with small heterogeneity present ($n = 6$; Hedges' $g = 0.20$, CI: [0.06, 0.35], $p = 0.006$; $Q = 4.56$, $p = .0.47$, $I^2 = 0.00\%$) and from baseline to latest follow-up (not including immediate post-intervention; $n = 12$; Hedges' $g = 0.19$, CI: [0.11, 0.28], $p < .001$; $Q = 14.11$, $p = 0.23$, $I^2 = 22.01\%$). See Figures 8 - 10 for forest plots of each of these time periods.

Behavior. The overall effect size for behavior from baseline through the latest follow-up (including post-intervention) was medium ($n = 13$; Hedges' $g = 0.48$, CI: [0.31, 0.66], $p < .001$). Heterogeneity among the included studies was high ($Q = 46.69$, $p < .001$, $I^2 = 74.30\%$). Significant changes were also observed from baseline to immediately post-intervention, with heterogeneity also significant and large ($n = 5$; Hedges' $g = 0.68$, CI: [0.23, 1.12], $p = 0.003$; $Q = 25.76$, $p < .001$, $I^2 = 84.47\%$), and from baseline to latest follow-up (not including immediate post-intervention; $n = 13$; Hedges' $g = 0.46$, CI: [0.29, 0.63], $p <$

.001; $Q = 44.39$, $p < .001$, $I^2 = 73.00\%$). See Figures 11 - 13 for forest plots of each of these time periods.

Distress. Additional analyses of changes in the trainees' distress level from baseline to latest follow-up revealed a small and marginally significant effect size ($n = 5$; Hedges' $g = 0.12$, CI: [-0.01, 0.24], $p = 0.08$). Heterogeneity for the included studies was small ($Q = 1.65$, $p = 0.80$, $I^2 = 0.00\%$). See Figure 14 for a forest plot of this outcome. Due to the limited number of studies reporting this outcome, we were only able to conduct analyses over the one time point.

Moderators

We examined several potential moderators to determine if other variables were contributing to the effect relationships and due to the presence of heterogeneity within our outcomes. Because we had studies rated as having low bias ($n = 8$) and others rated as having high bias ($n = 7$), we were interested in determining if the study's bias had an effect on the overall outcome. We calculated a continuous score of the total low bias ratings per item for each study (low = 1, uncertain = 0, high = 0) and conducted meta-regression analysis. The difference between studies based on bias and the overall effect of the three primary outcomes was significant ($B = -0.10$, CI: [-0.15, -0.03], $p = 0.003$). We followed up this significant effect with sensitivity analyses using only the studies rated as low bias. The overall effect size of combining the trainees' knowledge, attitude, and behavior was small (see Figure 15; Hedges' $g = 0.29$, CI: [0.16, 0.41], $p < .001$). The trainees' knowledge outcome had a moderate effect size (See Figure 16; Hedges' $g = 0.42$, CI: [0.24, 0.60], $p < .001$), while both the attitude and behavior outcomes had small effect sizes (see Figures 17 - 18; Hedges' $g =$

0.16, CI: [0.07, 0.24], $p < .001$ and Hedges' $g = 0.38$, CI: [0.19, 0.56], $p < .001$, respectively).

We also assessed differences between the nRCTs and the RCTs and found no moderating effect of study design (See Figure 19; $Q_B = 1.25$, $p = 0.26$). The non-randomized studies ($n = 4$) had a moderate effect size, Hedges' $g = 0.46$, CI: [0.22, 0.70], $p < .001$, and the randomized studies ($n = 11$) yielded a smaller but not significantly different effect, Hedges' $g = 0.31$, CI: [0.19, 0.42], $p < .001$. Notably, however, much more heterogeneity was present in the nRCTs, $Q = 6.30$, $p = 0.10$, $I^2 = 52.35\%$, than the RCTs, $Q = 12.45$, $p = 0.26$, $I^2 = 19.69\%$. In terms of investigator allegiance (see Figure 20; $Q_B = 0.003$, $p = 0.96$) and country of research (see Figure 21; $Q_B = 0.30$, $p = 0.59$), no significant moderation was present for either variable. Research teams not associated with the MHFA founders reported nearly identical effects as research teams involving the MHFA founders (Hedges' $g = 0.35$, CI: [0.23, 0.48], $p < .001$ vs. Hedges' $g = 0.34$, CI: [0.12, 0.56], $p = .003$). Studies conducted in Australia reported non-significantly different effects than all other countries (Hedges' $g = 0.39$, CI: [0.19, 0.58], $p < .001$ vs. Hedges' $g = 0.32$, CI: [0.19, 0.45], $p < .001$). We were unable to assess differences between studies published in the U.S. compared to other countries as originally planned due to a limited number of U.S. studies eligible for inclusion; however, we were able to make comparisons between Australia, where MHFA originated, and all other countries. Finally, gender was not found to have a moderating effect on the overall MHFA effect in trainees, $B = -0.002$, CI: [-0.009, 0.005], $p = 0.64$.

Publication Bias

Publication bias was assessed for both the overall effect as well as for each of the three main trainee outcomes. We were unable to assess publication bias with the recipient

data because three or more studies are needed to run the analyses. Trim-and-fill analyses of our overall effect indicated that five unpublished studies were likely not included. This adjusts the overall effect size from 0.34 to 0.27. See Figure 22 for a funnel plot before the adjustment and Figure 23 for a plot with the adjusted effect size, as well as a representation of where the five missing studies likely lie. Egger's regression intercept was 1.73, CI: [0.67, 2.78], $t = 3.53$, $p = 0.002$. This indicates that there is a statistically significant amount of asymmetry within these data. The fail-safe *N* was 251, indicating that quite a few null studies would need to have been missed before the cumulative effect would become statistically nonsignificant. Given the number of identified evaluation studies, it appears unlikely over 250 studies were missed. Thus, while the overall effect of MHFA may be slightly overstated it is unlikely that it is actually zero.

Knowledge. Trim-and-fill analyses of the overall trainee knowledge outcome indicated that two unpublished studies were likely not included, resulting in a slight adjustment in the effect size from 0.48 to 0.47 (see Figures 24 - 25). Egger's regression intercept was -0.10, CI: [-1.52, 1.33], $t = 0.15$, $p = 0.45$. This intercept value is close to zero meaning there is little asymmetry within this outcome. The fail-safe *N* was 449.

Attitude. Trim-and-fill analyses of the overall trainee attitude outcome indicated that five unpublished studies were likely not included, slightly reducing the overall effect size from 0.18 to 0.14 (see Figures 26 - 27). Egger's regression intercept was 1.42, CI: [0.18, 2.67], $t = 2.51$, $p = 0.01$. This is indicative of statistically significant asymmetry within this outcome. The fail-safe *N* was 86.

Behavior. Trim-and-fill analyses of the overall trainee behavior outcome indicated that six unpublished studies were likely not included. Notably, that brings the overall effect

size down from 0.48 to 0.26 (see Figures 28 - 29). Egger's regression intercept was 3.38, CI: [1.51, 5.25], $t = 4.00$, $p = 0.001$. This intercept value is above zero and is statistically significant, meaning these data are considerably asymmetrical. The fail-safe N was 324.

Recipient Effects

A limited number of studies reported on the recipient effects (those who are receiving aid on behalf of the trained individual) of MHFA. Two of our included studies (Jorm et al., 2010b; Lipson et al., 2014) examined recipient outcomes. Due to the scarcity of results in this area, we were unable to conduct many analyses for the outcomes (knowledge, personal stigma, help received or treatment sought, distress level), but some preliminary quantitative results are described along with a narrative review. Total effect sizes were calculated for changes between baseline and follow-up (six months for Jorm et al., 2010b, two to three months for Lipson et al., 2014). Overall, the findings between the quantitative and qualitative/narrative summary of the two studies suggests the impact on recipients may be more mixed.

Knowledge. There was no detected effect on the recipients' knowledge (see Figure 30; Hedges' $g = 0.02$, CI: [-0.06, 0.11], $p = 0.58$), though, no clear methods for increasing the recipients' knowledge were utilized in either study. Heterogeneity for the included studies was small ($Q = 0.004$, $p = 0.95$, $I^2 = 0.00\%$).

Attitude. No apparent effects were found for the recipients' personal stigma as well (see Figure 31; Hedges' $g = 0.04$, CI: [-0.005, 0.12], $p = 0.39$). Heterogeneity for the included studies was small ($Q = 0.003$, $p = 0.96$, $I^2 = 0.00\%$).

Help Received or Sought. Consistent with the preceding outcomes, recipients' help received or sought did not appear improved (see Figure 32; Hedges' $g = -0.04$, CI: [-0.026,

0.18], p =0.72). Heterogeneity for the included studies was small to moderate (Q = 1.57, p = 0.211, I^2 = 36.14%).

Distress. The recipients' distress level had a very small but non-significant effect (see Figure 33; Hedges' g = 0.12, CI: [-0.13, 0.38], p = 0.34). Heterogeneity for the included studies was moderate (Q = 2.74, p = 0.10, I^2 = 63.55%).

Recipient Narrative Review

Although the preliminary effect size estimates suggested no effect on the recipients, a closer examination of the two studies suggests mixed findings rather than no effect whatsoever. Jorm and colleagues (2010b) found limited effects on the recipients, who were students of trained teachers. They measured changes in the students' mental health knowledge, beliefs about seeking treatment, personal and perceived stigma, help received from a teacher, and personal mental health. No effects were apparent for the students. However, the authors did find that students of teachers who were trained in MHFA were more likely to report that they had received information about mental health problems from a teacher. Furthermore, on the post-test, these students were less likely to rate people with mental health issues as unpredictable (related to stigma).

Lipson and colleagues (2014) found no intervention effects for the residents of college campuses. They measured changes in the residents' mental health knowledge, personal stigma, personal mental health treatment, and distress. Notably, residents in the control group reported receiving more help than other students in the intervention group. In fact, Lipson and colleagues (2014) found that the intervention recipients had lower odds of reporting mental health problems, lower odds of seeking treatment, and lower odds of receiving informal support, with small effect sizes observed.

Discussion

Through a comprehensive search and meta-analytic methods, 15 MHFA trials were identified and systematically synthesized to evaluate the intervention's impact. To date, this is only the second meta-analysis conducted on MHFA, and the only review that relied exclusively on controlled trials. Though this somewhat limited our inclusion parameters, we believe that one can have more confidence in these findings and the true effects of the MHFA intervention program.

Overall, the results indicate that MHFA is an effective intervention (Hedges' g = 0.34) for trainees exposed to the curriculum. Yet, the effects of MHFA can be better understood through an examination of the primary outcomes of knowledge, attitude, and behavior and secondary outcome of distress. The primary outcome effect sizes for the trainees' changes in knowledge, attitude, and behavior from baseline to latest follow-up were in the small-to-moderate range. This is consistent with the previous meta-analysis (Hadlaczky et al., 2014), with which we had six overlapping included studies (Jorm et al., 2004; Jorm et al., 2010b; Kitchener & Jorm, 2004; Massey et al., 2014; O'Reilly, Bell, Kelly, & Chen, 2011; Svensson & Hansson, 2014). The largest effect sizes were found for the knowledge and behavior outcomes, with both being in the medium range, whereas the attitude effect size was within the small range. Primary outcome effect sizes somewhat varied between the current meta-analysis and Hadlaczky and colleagues' (2014) meta-analysis, though differing outcome time points in the included studies may have, in part, contributed to these differences.

The knowledge effect size (Hedges' g = 0.48) is slightly smaller than the previous meta-analysis (Glass's Δ = 0.56; Hadlaczky et al., 2014) but still indicates that the

intervention trainees increased in their mental health literacy as a result of being in the training. In addition to including whether the trainees correctly identified a disorder based on a vignette, we also included scores from a mental health literacy quiz, when available, rather than using the trainees' beliefs about treatments – as Hadlaczky and colleagues (2014) used, in addition to vignette data. We conceptualized the knowledge outcome as changes in literacy, identification, and available treatments as those seemed to be the most important components needed for the initial trainee and recipient identification and interaction.

In terms of impact on attitudinal changes, our results (Hedges' $g = 0.18$) suggest a smaller effect of MHFA that was also smaller than the previous meta-analysis (Glass's $\Delta = 0.28$; Hadlaczky et al., 2014). Reducing stigmatizing attitudes is not only important for the trainee and recipient interactions, but an adjustment in the trainees' beliefs may carryover to other contexts, such as communicating with others about mental illnesses. It was somewhat surprising that this effect was not greater, especially given Hadlaczky and colleagues' (2014) more robust finding. Due to the mixed literature regarding changes in perceived attitudes, we further assessed just the social distance and personal stigma components of this outcome. The pooled overall effect size for social distance and personal stigma scores was only slightly larger than with all three measures (Hedges' $g = 0.21$ vs. Hedges' $g = 0.18$). One possible explanation for this is a backfire effect, essentially meaning that as people learn more about stigma and social distance, they actually increase in negative attitudes rather than decrease (Levy & Maaravi, 2017). Future research should examine these differences more closely to assess whether perceived stigma is an adequate measure of attitudinal change within MHFA.

Our analysis yielded an effect size for the behavioral outcome (Hedges' $g = 0.48$) on par with the effect for the knowledge outcome. Notably, this effect is more pronounced than

the previous meta-analysis's finding (Glass's Δ = 0.25; Hadlaczky et al., 2014). A plausible contributing factor for this large increase in effect is our use of trainees' self-reported confidence in providing help within this outcome, in addition to actual help provided. Although the overall effect size for confidence (Hedges' $g = 0.52$; $n = 12$) was greater than actual help provided (Hedges' $g = 0.35$; $n = 7$), it seemed like a necessary intermediary measure between changes in knowledge and attitudes to actually providing the aid. Research involving the Theory of Planned Behavior (TPB) suggests that confidence is one of the strongest predictors of behavior change (Ajzen, 1991). Furthermore, the number of studies reporting actual help provided was smaller than those reporting confidence ratings; therefore, combining the two measures increased the total sample size for this outcome. Even though our effect size for this outcome was moderate in magnitude, an analysis of the heterogeneity revealed significant variability among the studies. The behavior outcome has consistently been one that researchers have had the most difficulty assessing and therefore, have found more mixed results. Thus, further investigation into conceptualizing and measuring this outcome seems warranted.

In terms of the secondary outcome of the trainees' psychological distress, the impact of MHFA was found to have a small and marginally significant effect size (Hedges' $g = 0.12$). Notably, only about a third of our included studies measured changes in the trainees' psychological distress, and, on average, the trainees' total distress levels were within the average range. Although MHFA is designed to help improve the well-being of the recipients, the trainees may also be able to apply some of the course content to themselves and their own mental health awareness and care (Davies et al., 2018). Furthermore, this is the first known

synthesis of this outcome within the MHFA and gatekeeper literature. Results indicate that this outcome may be affected by MHFA and thus, could benefit from further evaluation.

To some extent heterogeneity was present in all of the outcomes, with the behavior outcome being notably larger. However, the significant risk of bias moderator provides further rationale for the modest heterogeneity present in the included studies, especially given that we rated about half (n = 7) of the included studies as highly biased. Nonetheless, it is important to note that although many of the studies assessed the three outcomes, they were not uniformly measured across all of our included studies. Furthermore, in some cases, there was limited consistency of outcome measure selection and conceptualization, which restricted the data we were able to include and thus, may limit the conclusions that can be drawn. Future research would benefit from clearer conceptualizations of knowledge, attitude, and behavior as well as greater consensus in regard to how to measure each of these domains. In some cases where population-specific measures are preferable (e.g., O'Reilly et al., 2011), it seems reasonable to consider creating both general and specific outcome measures, to be used in any context.

The only significant moderator, identified a priori, was with study quality. We looked at this moderator as a continuous variable (total number of low bias items for each study) rather than a total dichotomous (low vs. high) one, to allow for a more precise estimate of this effect. We found that the bias level of the studies had a significant effect on the outcomes (B = -0.10, p = 0.003). After removing high bias studies, the effect sizes for the three primary outcomes (knowledge g = 0.42 vs. 0.48; attitude g = 0.15 vs. 0.18; behavior g = 0.38 vs. 0.48), as well as the overall effect of all three outcomes combined (g = 0.29 vs. 0.34), were consistently smaller, yet still statistically significant. Research has shown that studies with a

high risk of bias are more likely to produce over-inflated effect sizes due to the employment of less rigorous study methodology (Schulz, Altman, Moher, & CONSORT Group, 2010). Thus, future research should continue to focus on minimizing sources of bias (e.g., randomization, allocation concealment, incomplete outcome data).

As far as the other potential moderators, the differences between sub-groups were small to indistinguishable and, therefore, not statistically significant. Although RCTs are often considered the gold standard for intervention studies because they allow researchers to establish efficacy and make causal inferences about the effects of treatments, we found that they only had a slightly smaller effect size in relation to the nRCTs (Hedges' $g = 0.31$ and 0.46, respectively) and that randomization status, in and of itself, did not significantly affect the outcomes. Surprising to us, the potential moderating effect of the investigator allegiance was also not significant. The overall outcome difference between research teams who were unaffiliated with the MHFA developers compared to research teams involving the founders, was statistically indistinguishable (Hedges' $g = 0.35$ and 0.34, respectively). This indicates that the MHFA founders have not found greater or more positive results than other researchers, which is notable as this has historically drawn criticism. We also assessed the effect of country of research. Although we had originally hoped to compare studies conducted in the U.S. to studies in other countries, there were insufficient numbers of U.S. studies to meaningfully do so. Therefore, we pooled all of the studies that were conducted in countries other than Australia and compared those to the studies conducted only in Australia. Non-Australian based research teams found slightly smaller effects than Australia-based teams; however, this difference was not statistically significant. Finally, we did not find any moderating effects of gender. The only study that examined the potential differential effects

of gender was Lipson and colleagues (2014) who found that female RAs (trainees) had less contact with the residents (recipients) about their mental health and thus, experienced less of an intervention effect. Additional post-hoc analyses revealed that the gender of the trainee did not clearly influence changes in the three primary outcomes for trainees, though there was a marginally significant ($p < .10$) effect with knowledge. Given gender differences observed in the prevalence of some disorders as well as treatment seeking behavior (Eaton et al., 2012; Wendt & Shafer, 2016), future investigators may want to more carefully consider this potential factor on effects for trainees and recipients.

Unlike the previous meta-analytic review (Hadlaczky et al., 2014), we examined changes at differing time points (i.e., baseline to immediate post and baseline to latest follow-up). Post-intervention measures for the included studies varied from immediately after the intervention to weeks or months later. Primary outcome effect sizes for baseline to immediately post-intervention were consistently greater than for baseline to latest follow-up (knowledge $g = 0.56$ vs. 0.50; attitude $g = 0.20$ vs. 0.19; behavior $g = 0.68$ vs. 0.46). For the knowledge and attitude outcomes, it seems likely that the effect sizes would be larger immediately after the training rather than weeks or months later because the course material would be more recent for the trainees. Nonetheless, it appears that these two outcomes are rather durable as they do not vary much between time points.

On the other hand, one would expect that effect sizes for the behavior outcome would increase over time as trainees have more time (and thus, opportunities) to actually provide aid, but this was not observed (Hedges' $g = 0.68$ vs. 0.46). The follow-up periods were relatively short, and trainees may not have had many opportunities to offer help, which may be a reason that the follow-up effect size was smaller than the immediate post-intervention

effect size. However, one way to more indirectly measure helping behavior is through assessing the trainees' help-seeking intentions, though this was not included as one of the behavior measures in the current meta-analysis due to the limited number of studies reporting it (Burns et al., 2017; Davies et al., 2018; Jorm et al., 2004). Help-seeking intention is measured through trainees' ratings towards how they would respond to the individual described in the vignette; correct responses are scored based on the trainees' mention of the components of the MHFA *Action Plan.*

In relation to all three primary outcomes, the length of the follow-up period was noted as a limitation or future direction in many of our included studies (Burns et al., 2017; Jorm et al., 2010b; Lipson et al., 2014; Moll, Patten, Stuart, MacDermid, & Kirsh, in press; O'Reilly et al., 2011; Wong, Lau, Kwok, Wong, & Tori, 2017). This meta-analysis's findings, as well as those from studies that measured outcomes both immediately after the intervention and at a later follow-up point, question whether the need for additional booster sessions may be needed to remind the trainees about the key components of MHFA. Notably, one study (Svensson & Hansson, 2014) included a brief two-year follow-up, though they only assessed changes for the intervention group, not the control group. Hence, examining the enduring effects of MHFA seems to be an important area for further research.

Generally, MHFA research has been based on self-report measures and perceptions of changes in oneself, rather than objective measures. Socially desirable responding and/or only assessing the trainees' perceived changes in each outcome were noted as limitations in some studies (Jorm et al., 2010a, 2010b; Wong et al., 2017). Interestingly, three of the included studies suggested involving people who have a mental health disorder or have experienced a mental health crisis (i.e., "mental health ambassador;" Jensen et al., 2011) in the actual

training (Jensen et al., 2016; Jorm et al., 2010a; O'Reilly et al., 2011). Previous research has found that increased contact with people who have mental illnesses has positive effects on stigma (Corrigan & Penn, 1999; Jorm & Oh, 2009; O'Reilly, Bell, & Chen, 2010). It would be interesting to see if this contact could also increase the trainees' confidence and actual helping behavior. Furthermore, this sort of direct experience could provide trainees with more realistic opportunities to practice their new skills, as well as to debrief about their experiences with the MHFA trainer(s) and other trainees (Lipson et al., 2014). Moreover, this may provide an opportunity to observe and measure the trainees' skill use and recipient effects more directly with less reliance on self-report measures.

Finally, and of critical importance, we conducted preliminary analyses to assess the effects for the recipients of the aid. To the best of our knowledge, this is the first effort to synthesize MHFA effects on recipients. Our analyses were limited in scope because of the small number of controlled studies that have evaluated such effects (Jorm et al., 2010b; Lipson et al., 2014) and therefore should be considered preliminary. That said, we were unable to detect effects for the recipients in any of the three primary outcomes: knowledge (Hedges' $g = 0.02$), attitude (Hedges' $g = 0.04$), and treatment sought or received (Hedges' $g = -0.04$), nor for psychological distress. Recipients did not receive any formal intervention directed at affect their knowledge or attitudes; however, both study authors noted that they had hoped to see these improve as a result of increased interactions with the trainees. With regards to the behavior outcome, it is based upon potential recipients' self-reported contact with a trainee. The nature of this type of reporting appears somewhat problematic. Furthermore, measuring the recipients' help-seeking behavior (e.g., changes in referrals to the campus counseling center) could be indicative of a MHFA trainee encouraging a

recipient to seek services, but one cannot draw a clear relationship there. Also, a referral to professional treatment is not always needed or the most appropriate way of intervening (Lipson et al., 2014). Finally, as noted with the trainee outcomes, the relatively short follow-up periods may not have given the trainees sufficient opportunities to put their skills to use. For example, trained teachers in Jorm and colleagues' (2010b) study could have provided aid to the next year's students, but this would not have been captured during the study's follow-up period of six months. Given the limited research in this area, our conclusions are speculative but offer important considerations for future research in this area.

With only two studies included in the analyses, we lacked adequate power to detect effects; however, our findings are generally consistent with previous research that has attempted to assess recipient effects (e.g., Booth et al., 2017). This points to major limitations within the current literature on MHFA: the scarcity of interventions that assess the effects on the recipients of the aid. As Watts (2017) recently pointed out, the positive MHFA trainee effects can be viewed similarly to instructors teaching children about math and then reporting that they are more confident performing the computations while failing to assess differences for the children. Granted, assessing MHFA recipient effects may not be quite this straightforward. It seems entirely plausible that one of the main factors contributing to the dearth of recipient outcome literature is the feasibility of collecting this data. Both Jorm and colleagues (2010b) and Lipson and colleagues (2014) suggested the following factor as one possible reason for their limited recipient effects. Research (e.g., Jorm et al., 2010b; Lipson et al., 2014; Pierce et al., 2010) has found that, at least in academic and sports contexts, the recipients were more likely to report that they would seek help from a family member rather than a teacher or coach. Therefore, expanding MHFA trainings to parents could help capture

more direct contact between trainees and recipients, though this primarily pertains to youth-focused settings.

It is important to remember that the target trainee population for MHFA is the general public, meaning the target recipient population is also the general population (Kitchener & Jorm, 2002a, 2002b). The practicality of collecting data from even a sample of the general population seems quite challenging. Only a few settings come to mind that allow for some control over the intervention recipient population and capturing of utilization of services data (e.g., academic setting: Jorm et al., 2010b; Lipson et al., 2014; sports setting: Pierce et al., 2010; military setting: Mohatt, Boeckmann, Winkel, Mohatt, & Shore, 2017). Notably, these settings have both diversity of personnel as well as a breadth of supports and services available within them. These types of settings allow for more control over who is within the sample, yet still include representative groups of people. Thus, it seems logical to continue utilizing these types of settings as well as explore others to further assess recipient effects.

Overall, the results of this meta-analysis support the effectiveness of MHFA for the trainees yet are inconclusive for the recipients. Therefore, future research assessing recipient effects clearly seems warranted. Furthermore, future research should continue to employ rigorous controlled designs to test future MHFA adaptations. Recent research has examined the effects of alternative MHFA interventions (i.e., e-based learning; Davies et al., 2018) as well as a comparison of MHFA to other active treatments with a control group (i.e., Beyond Silence; Moll et al., in press). Both of these appear to be interesting directions to pursue in future research and to further demonstrate MHFA's effectiveness. At this point, research has consistently shown that MHFA is an effective intervention for trainees, which is important.

However, if the intervention is not affecting recipients as originally intended, then programmatic adjustments may need to be made.

References

Aakre, J. M., Lucksted, A., & Browning-McNee, L. (2016). Evaluation of youth Mental Health First Aid USA: A program to assist young people in psychological distress. *Psychological Services, 13*(2), 121-126. doi:10.1037/ser0000063

Ajzen, I. (1991) The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211. doi:10.1016/0749-5978(91)90020-T

Booth, A., Scantlebury, A., Hughes-Morley, A., Mitchell, N., Wright, K., Scott, W., & McDaid, C. (2017). Mental health training programmes for non-mental health trained professionals coming into contact with people with mental ill health: A systematic review of effectiveness. *BMC Psychiatry*, *17*(196). doi:10.1186/s12888-017-1356-5

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. Chichester, U.K.: John Wiley & Sons. doi:10.1002/9780470743386

*Burns, S., Crawford, G., Hallett, J., Hunt, K., Chih, H. J., & Tilley, P. M. (2017). What's wrong with John? A randomised controlled trial of Mental Health First Aid (MHFA) training with nursing students. *BMC Psychiatry, 17*(111). doi:10.1186/s12888-017-1278-2

Cheng, H., Wang, C., McDermott, R. C., Kridel, M., & Rislin, J. L. (2018). Self-stigma, mental health literacy, and attitudes toward seeking psychological help. *Journal of Counseling & Development, 96*(1), 64-74. doi:10.1002/jcad.12178

Corrigan, P. W., & Penn, D. L. (1999). Lessons from social psychology on discrediting psychiatric stigma. *American Psychologist*, *54*(9), 765-776.

*Davies, E. B., Beever, E., & Glazebrook, C. (2018). A pilot randomized controlled study of the Mental Health First Aid eLearning course with UK medical students. *BMC Medical Education*, *18*(45). doi:10.1186/s12909-018-1154-x

Dahlberg, K. M., Waern, M., & Runeson, B. (2008). Mental health literacy and attitudes in a Swedish community sample: Investigating the role of personal experience of mental health care. *BMC Public Health, 8*(8). doi:10.1186/1471-2458-8-8

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463. doi:10.1111/j.0006-341x.2000.00455.x

Eaton, N. R., Keyes, K. M., Krueger, R. F., Balsis, S., Skodol, A. E., Markon, K. E., . . . Hasin, D. S. (2012). An invariant dimensional liability model of gender differences in mental disorder prevalence: Evidence from a national sample. *Journal of Abnormal Psychology, 121*(1), 282-288. doi:10.1037/a0024780

Feldman, D. B., & Crandall, C. S. (2007). Dimensions of mental illness stigma: What about mental illness causes social rejection? *Journal of Social and Clinical Psychology, 26*(2), 137-154. doi:10.1521/jscp.2007.26.2.137

Graham, R. (2015, January 23). The promise and limits of 'mental health first aid'. Retrieved from https://www.bostonglobe.com/ideas/2015/01/23/the-promise-and-limits-mental-health-first-aid/TaSVSJma5TMnHZZzNItYKM/story.html

Griffiths, K. M., Christensen, H., Jorm, A. F., Evans, K., & Groves, C. (2004). Effect of web based depression literacy and cognitive–behavioural therapy interventions on stigmatising attitudes to depression. *British Journal of Psychiatry, 185*(4), 342-349. doi:10.1192/bjp.185.4.342

Gryglewicz, K., Childs, K. K., & Soderstrom, M. F. (2018). An evaluation of Youth Mental Health First Aid training in school settings. *School Mental Health, 10*(1), 48-60. doi:10.1007/s12310-018-9246-7

Hadlaczky, G., Hökby, S., Mkrtchian, A., Carli, V., & Wasserman, D. (2014). Mental Health First Aid is an effective public health intervention for improving knowledge, attitudes, and behaviour: A meta-analysis. *International Review of Psychiatry, 26*(4), 467-475. doi:10.3109/09540261.2014.924910

Hedges L. V., & Olkin I. (1985). Statistical methods for meta-analysis. San Diego, CA: Academic Press.

Higgins, J. P., & Altman, D. G. (2008). Assessing risk of bias in included studies. In *Cochrane handbook for systematic reviews of interventions* (pp. 187-241). Wiley. doi:10.1002/9780470712184.ch8

Higgins, J. P., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., . . . Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal, 343.* doi:10.1136/bmjd5928

Insel, T. R. (2008). Assessing the economic costs of serious mental illness. *American Journal of Psychiatry*, *165*(6), 663-665. doi:10.1176/appi.ajp.2008.08030366

*Jensen, K. B., Morthorst, B. R., Vendsborg, P. B., Hjorthøj, C., & Nordentoft, M. (2016). Effectiveness of Mental Health First Aid training in Denmark: A randomized trial in waitlist design. *Social Psychiatry and Psychiatric Epidemiology, 51*(4), 597-606. doi: 10.1007/s00127-016-1176-9

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80*(1), 94-106. doi:10.1037/00219010.80.1.94

*Jorm, A. F., Kitchener, B. A., O'Kearney, R., & Dear, K. B. G. (2004). Mental Health First Aid training of the public in a rural area: A cluster randomized trial. *BMC Psychiatry, 4*(33). doi:10.1186/1471-244X-4-33

Jorm, A. F., Nakane, Y., Christensen, H., Yoshioka, K., Griffiths, K. M., & Wata, Y. (2005). Public beliefs about treatment and outcome of mental disorders: A comparison of Australia and Japan. *BMC Medicine, 3*(12). doi:10.1186/1741-7015-3-12

*Jorm, A. F., Kitchener, B. A., Fischer, J. A., & Cvetkovski, S. (2010a). Mental Health First Aid training by e-learning: A randomized controlled trial. *Australian and New Zealand Journal of Psychiatry, 44*(12)*,* 1072-1081. doi:10.3109/00048674.2010.516426

*Jorm, A. F., Kitchener, B. A., Sawyer, M. G., Scales, H., & Cvetkovski, S. (2010b). Mental Health First Aid training for high school teachers: A cluster randomized trial. *BMC Psychiatry*, *10*(51). doi:10.1186/1471-244X-10-51

Jorm, A. F., & Kitchener, B. A. (2011). Noting a landmark achievement: Mental Health First Aid training reaches 1% of Australian adults. *Australian and New Zealand Journal of Psychiatry, 45,* 808-813. doi:10.3109/00048674.2011.594785

Jorm, A. F., & Oh, E. (2009). Desire for social distance from people with mental disorders: A review. *Australian & New Zealand Journal of Psychiatry*, *43*(3) 183-200. doi: 10.1080/00048670802653349.

Jung, H., Sternberg, K. V., & Davis, K. (2017). The impact of mental health literacy, stigma, and social support on attitudes toward mental health help-seeking. *International Journal of Mental Health Promotion, 19*(5), 252-267. doi:10.1080/14623730.2017.1345687

Kelly, C., Jorm, A. F., & Wright, A. (2007). Improving mental health literacy as a strategy to facilitate early intervention for mental disorders. *Medical Journal of Australia*, *187*(7), S26-S30.

Kelly, C. M., Mithen, J. M., Fischer, J. A., Kitchener, B. A., Jorm, A. F., Lowe, A., & Scanlan, C. (2011). Youth Mental Health First Aid: A description of the program and an initial evaluation. *International Journal of Mental Health Systems*, *5*(4). doi:10.1186/1752-4458-5-4

Kessler, R., Andrews, G., Colpe, L., Hiripi, E., Mroczek, D., Normand, S., . . . Zaslavsky, A. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine, 32*(6), 959-976. doi:10.1017/s0033291702006074

Kitchener, B. A., & Jorm, A. F. (2002a). *Mental health first aid manual*. Canberra: Center for Mental Health Research, Australian National University.

Kitchener, B. A., & Jorm, A. F. (2002b). Mental Health First Aid training for the public: Evaluation of effects on knowledge, attitudes, and helping behavior. *BMC Psychiatry, 2*(10). doi:10.1186/1471-244X-2-10

*Kitchener, B. A., & Jorm, A. F. (2004). Mental Health First Aid training in a workplace setting: A randomized controlled trial. *BMC Psychiatry, 4*(23). doi:10.1186/1471-244X-4-23

Kitchener, B. A., & Jorm, A. F. (2006). Mental Health First Aid training: Review of evaluation studies. *Australian and New Zealand Journal of Psychiatry, 40*(1), 6-8. doi:10.1080/j.1440-1614.2006.01735.x

Levy, A., & Maaravi, Y. (2017). The boomerang effect of psychological interventions. *Social Influence, 13*(1), 39-51. doi:10.1080/15534510.2017.1421571

Link, B. G., Phelan, J. C., Bresnahan, M., Stueve, A., & Pescosolido, B. A. (1999). Public conceptions of mental illness: Labels, causes, dangerousness, and social distance. *American Journal of Public Health, 89*(9), 1328-1333. doi:10.2105/ajph.89.9.1328

*Lipson, S. K., Speer, N., Brunwasser, S., Hahn, E., & Eisenberg, D. (2014). Gatekeeper training and access to mental health care at universities and colleges. *Journal of Adolescent Health, 55*(5), 612-619. doi:10.1016/j.jadohealth.2014.05.009

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the depression anxiety stress scales* (2nd ed.). Sydney: Psychology Foundation Monograph.

*Massey, J., Brooks, M., & Burrow, J. (2014). Evaluating the effectiveness of Mental Health First Aid training among student affairs staff at a Canadian university. *Journal of Student Affairs Research and Practice, 51*(3), 323-336. doi:10.1515/jsarp-2014-0032

*Moffitt, J., Bostock, J., & Cave, A. (2014). Promoting well-being and reducing stigma about mental health in the fire service. *Journal of Public Mental Health, 13*(2), 103-113. doi:10.1108/JPMH-02-2013-0004

*Mohatt, N. V., Boeckmann, R., Winkel, N., Mohatt, D., & Shore, J. (2017). Military Mental Health First Aid: Development and preliminary efficacy of a community training for

improving knowledge, attitudes, and helping behaviors. *Military Medicine, 182*(1), e1576-e1583.

Moher D., Liberati A., Tetzlaff J., Altman D. G., & The PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7). doi:10.1371/journal.pmed.1000097

*Moll, S., Patten, S. B., Stuart, H., MacDermid, J. C., & Kirsh, B. (in press). Beyond silence: A randomized, parallel-group trial exploring the impact of workplace mental health literacy training with healthcare employees. *Canadian Journal of Psychiatry*. doi:10.1177/0706743718766051

Morgan, A. J., Reavley, N. J., Jorm, A. J., & Beatson, R. (2017). Discrimination and support from friends and family members experienced by people with mental health problems: Findings from an Australian national survey. *Social Psychiatry and Psychiatric Epidemiology, 52*(11), 1395-1403. doi:10.1007/s00127-017-1391-z

National Academies of Sciences, Engineering, and Medicine. (2016). Ending discrimination against people with mental and substance use disorders: The evidence for stigma change. Washington, DC: The National Academies Press. doi:10.17226/23442

National Council for Behavioral Health. (2014). *Mental Health First Aid State Policy Toolkit*. Retrieved February 9, 2017, from http://www.thenationalcouncil.org/wp-content/uploads/2014/08/Policy-Toolkit-FINAL.pdf.

O'Reilly, C. L., Bell, J. S., & Chen, T. F. (2010). Consumer-led mental health education for pharmacy students: Impact on stigma and attitudes towards provision of pharmaceutical services. *American Journal of Pharmaceutical Education, 74*(9), 167.

*O'Reilly, C. L., Bell, J. S., Kelly, P. J., & Chen, T. F. (2011). Impact of Mental Health First Aid training on pharmacy students' knowledge, attitudes and self-reported behaviour: A controlled trial. *Australian and New Zealand Journal of Psychiatry, 45*(7), 549-557. doi:10.3109/00048674.2011.585454

Parcesepe, A. M., & Cabassa, L. J. (2013). Public stigma of mental illness in the United States: A systematic literature review. *Administration and Policy in Mental Health and Mental Health Services Research, 40*(5), 384-399. doi:10.1007/s10488-012-0430-z

Pescosolido, B. A., Jensen, P. S., Martin, J. K., Perry, B. L., Olafsdottir, S., & Fettes, D. (2008). Public knowledge and assessment of child mental health problems: Findings from the National Stigma Study - Children. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(3)*,* 339-349. doi:10.1097/CHI.0b013e318160e3a0

Pierce, D., Liaw, S., Dobell, J., & Anderson, R. (2010). Australian rural football club leaders as mental health advocates: An investigation of the impact of the coach the coach project. *International Journal of Mental Health Systems, 4*(10). doi:10.1186/1752-4458-4-10

Rosenberg, L., & Germanotta, C. (2018, February 19). Lady Gaga's commitment to mental health training sets the standard for everyone. Retrieved from http://www.philly.com/philly/opinion/commentary/mental-health-united-states-lady-gaga-trainings-opinion-20180219.html

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. doi:10.1037/0033-2909.86.3.638

Rothstein, H. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester: Wiley.

SAMHSA. (2012, May). Intervention summary - Mental Health First Aid. Retrieved May 8, 2018, from https://www.nrepp.samhsa.gov/Legacy/ViewIntervention.aspx?id=321

Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *Annals of Internal Medicine*, *152*, 726-732. doi:10.7326/0003-4819-152-11-201006010-00232

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., & Silove, D. (2014). The global prevalence of common mental disorders: A systematic review and meta-analysis 1980-2013. *International Journal of Epidemiology, 43*(2), 476-493. doi:10.1093/ije/dyu038

*Svensson, B., & Hansson, L. (2014). Effectiveness of Mental Health First Aid training in Sweden. A randomized controlled trial with a six-month and two-year follow-up. *PLoS ONE*, *9*(6). doi:10.1371/journal.pone.0100911

Swarbrick, P., & Brown, J. (2015). Forward. In *Mental health first aid USA (*1st ed., revised, pp. ii-vii). Washington DC: National Council for Behavioral Health.

Wang, J., Adair, C., Fick, G., Lai, D., Evans, B., Perry, B. W., . . . Addington, D. (2007). Depression literacy in Alberta: Findings from a general population sample. *Canadian Journal of Psychiatry, 52*(7)*,* 442-449. doi:10.1177/070674370705200706

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., … Wells, J. E. (2007). Worldwide use of mental health services for anxiety, mood,

and substance disorders: Results from 17 countries in the WHO World Mental Health (WMH) surveys. *Lancet*, *370*(9590), 841-850. doi:10.1016/S0140-6736(07)61414-7

Ware, J. E., Kosinsk, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*(3), 220-233. doi:10.1097/00005650-199603000-00003

Watts, J. (2017, July 15). The mental health first aid programme is a pet project – if the NHS services were properly funded in the first place, it wouldn't be needed. Retrieved from https://www.independent.co.uk/voices/mental-health-first-aid-theresa-may-depression-anxiety-nhs-underfunded-services-turned-away-a7842571.html

Wendt, D., & Shafer, K. (2016). Gender and attitudes about mental health help seeking: Results from national data. *Health & Social Work*, *41*(1), e20-e28. doi:10.1093/hsw/hlv089

Whiteford, H., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., ... Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *Lancet*, *382*(9904), 1575-1586. doi:10.1016/S0140-6736(13)61611-6

White House, President Barack Obama. (2013). Now is the time: The President's plan to protect our children and our communities by reducing gun violence. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/docs/wh_now_is_the_time_full.pdf

*Wong, D. F. K., Lau, Y., Kwok, S., Wong, P., & Tori, C. (2017). Evaluating the effectiveness of Mental Health First Aid program for Chinese people of Hong Kong. *Research on Social Work Practice, 27*(1), 59-67. doi:10.1177/1049731515585149

Wong, E. C., Collins, R. L., & Cerully, J. L. (2015). *Reviewing the evidence base for Mental Health First Aid: Is there support for its use with key target populations in California?* Santa Monica, CA: Rand Corporation. Retrieved from http://www.rand.org/pubs/research_reports/RR972.html.

Table 1

Cohen's Kappa Agreement for Coders

	Investigator 1 include	Investigator 2 exclude	Total
Investigator 1 include	20	10	30
Investigator 2 exclude	7	6,240	6,247
Total	27	6,250	6,277

Overall Cohen's kappa = .70

Overall rating agreement = 99.7%

N = 6,277

Table 2

Summary of Included Studies

Study	MHFA format	Hours	Time points	Outcomes	Study location	Control condition	*n*	Female (%)	Age (*M*)	Bias rating
				Randomized Control Trials						
Burns et al. (2017)	Adult; F2F	13	Pre, post, f/u-2 mos	K: quiz; vignette A: personal, perceived stigma; SD B: confidence	Australia	Comparison	181	65	24.1	High
Davies et al. (2018)	Adult; eL	6-8	Pre, f/u-6 wks	A: personal stigma B: confidence; actual help	U.K.	Comparison	55	65	19.9	Low
Jensen et al. (2016)	Adult; F2F	12	Pre, f/u-6 mos	K: quiz A: personal, perceived stigma; SD B: confidence; actual help	Denmark	Comparison	560	84	43.0	High
Jorm et al. (2004)	Adult; F2F	9	Pre, f/u-4 mos	K: vignettes A: SD B: confidence; actual help	Australia	Waitlist	753	52	47.6	Low

Jorm et al., (2010a)	Youth; F2F	14	Pre, post, f/u-6 mos	K: quiz; vignette A: personal, perceived stigma B: confidence; actual help	Australia	Waitlist	1,960	56	--	Low
	Recipients			K: vignette A: personal stigma B: help received			1,633	54	13.7	--
Jorm et al. (2010b)	Adult; eL	--	Pre, post, f/u-6 mos	K: vignettes A: personal, perceived stigma; SD B: confidence; actual help	Australia	Waitlist	262	81	40	Low
Kitchener & Jorm (2004)	Adult; F2F	9	Pre, f/u-5 mos	K: vignettes A: SD B: confidence; actual help	Australia	Waitlist	301	78	39.4	Low
Lipson et al. (2014)	Adult; F2F	12	Pre, f/u-2 mos	K: quiz A: personal stigma B: confidence	U.S.A.	Comparison	2,543	49	20.4	Low
	Recipients			K: quiz A: personal stigma B: help received			1,990	63.9	19.1	--

Mohatt et al. (2017)	Adult; F2F	8	Pre, post, f/u-4 mos, f/u-8 mos	K: quiz B: actual help	U.S.A.	Comparison	176	--	--	High
Moll et al. (in press)	Adult; F2F	12	Pre, post, f/u-3 mos	K: quiz A: SD B: confidence; actual help	Canada	Active treatment	192	89	44.7	Low
Svensson & Hansson (2014)	Adult; F2F	12	Pre, f/u-2 mos	K: quiz A: personal, perceived stigma; SD B: confidence; actual help	Sweden	Comparison	406	77	45.6	Low
non-Randomized Control Trials										
Massey, Brooks, & Burrow (2014)	Adult; F2F	12	Pre, f/u-6 mos	K: quiz B: confidence	Canada	Comparison	84	--	--	High
Moffitt, Bostock, & Cave (2014)	Adult; F2F	12	Pre, post	K: quiz A: perceived stigma; SD	U.K.	Leaflets	176	23	43*	High
O'Reilly et al. (2011)	Adult; F2F	12	Pre, f/u-6 wks	K: vignettes A: SD	Australia	Comparison	258	64	21	High
Wong et al. (2017)	Adult; F2F	12	Pre, f/u-6 wks	K: vignettes A: personal, perceived stigma; SD B: confidence	Hong Kong	Seminars	277	68	34.9	High

Note. F2F = face-to-face; eL = e-learning; f/u = follow-up; K = knowledge; A = attitude; B = behavior; SD = social distance; *modal age

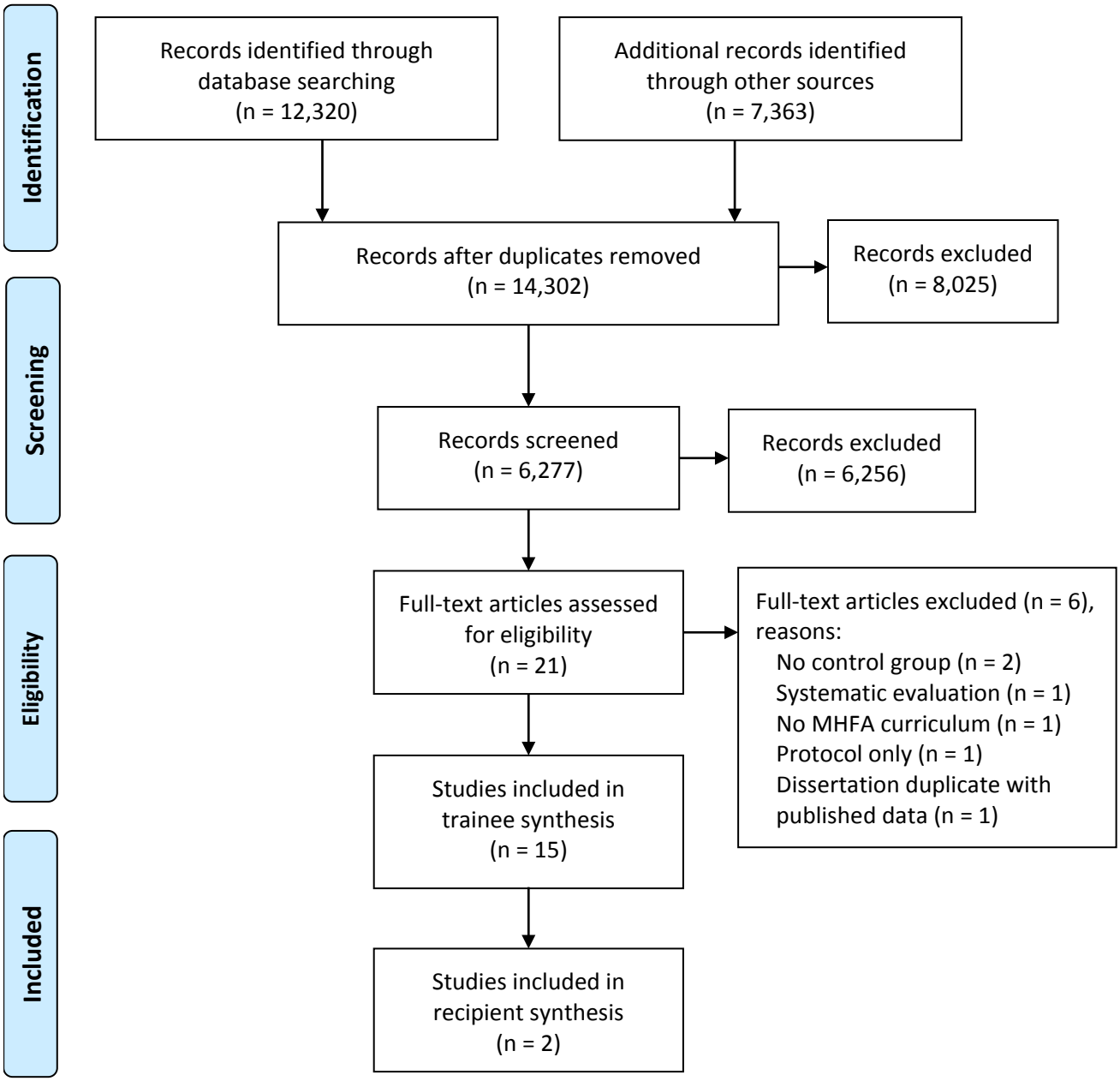
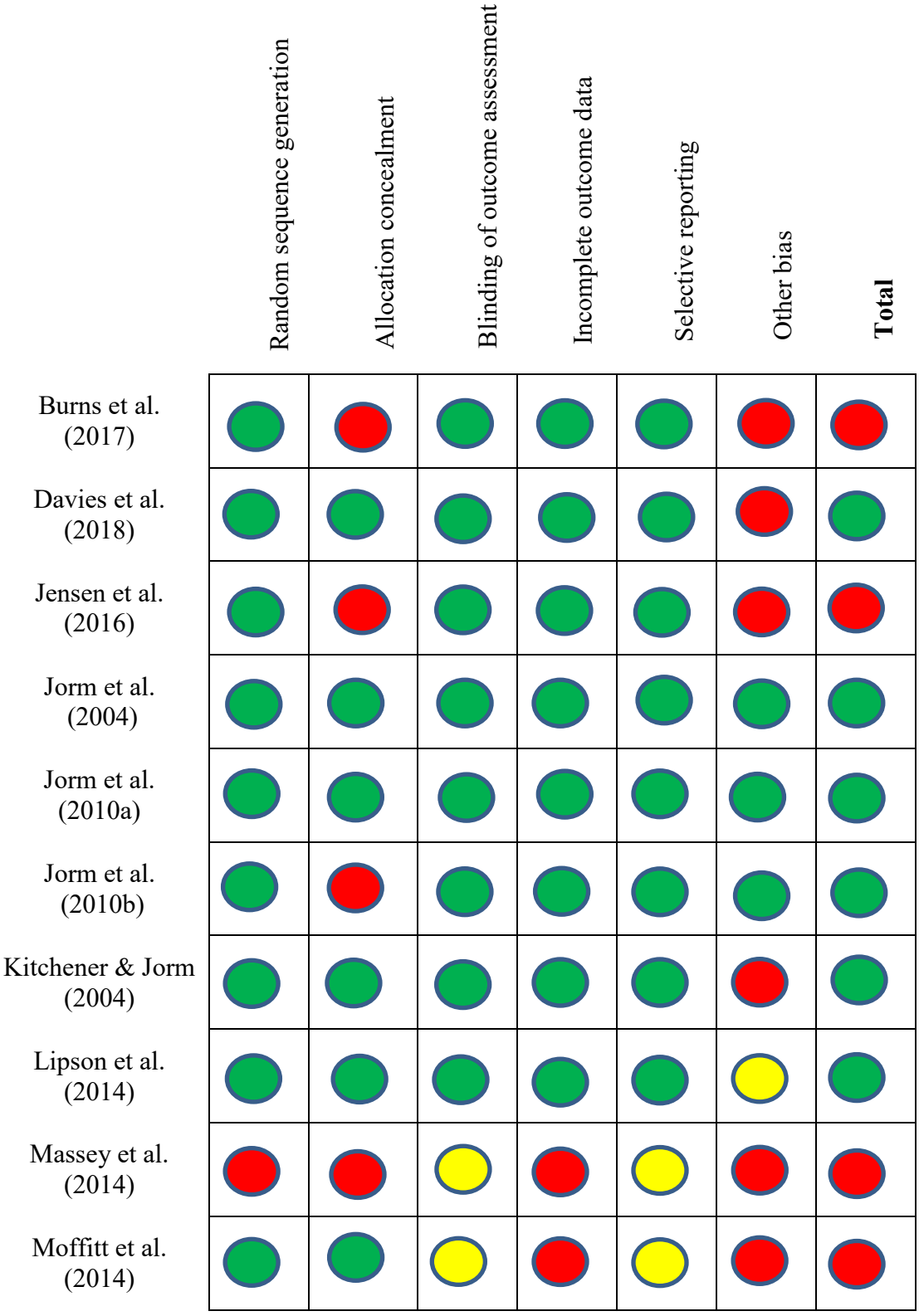


Figure 1. PRISMA flow chart.


Random sequence generation
Allocation concealment
Blinding of outcome assessment
Incomplete outcome data
Selective reporting
Other bias
Total
Burns et al. (2017)
Davies et al. (2018)
Jensen et al. (2016)
Jorm et al. (2004)
Jorm et al. (2010a)
Jorm et al. (2010b)
Kitchener & Jorm (2004)
Lipson et al. (2014)
Massey et al. (2014)
Moffitt et al. (2014)

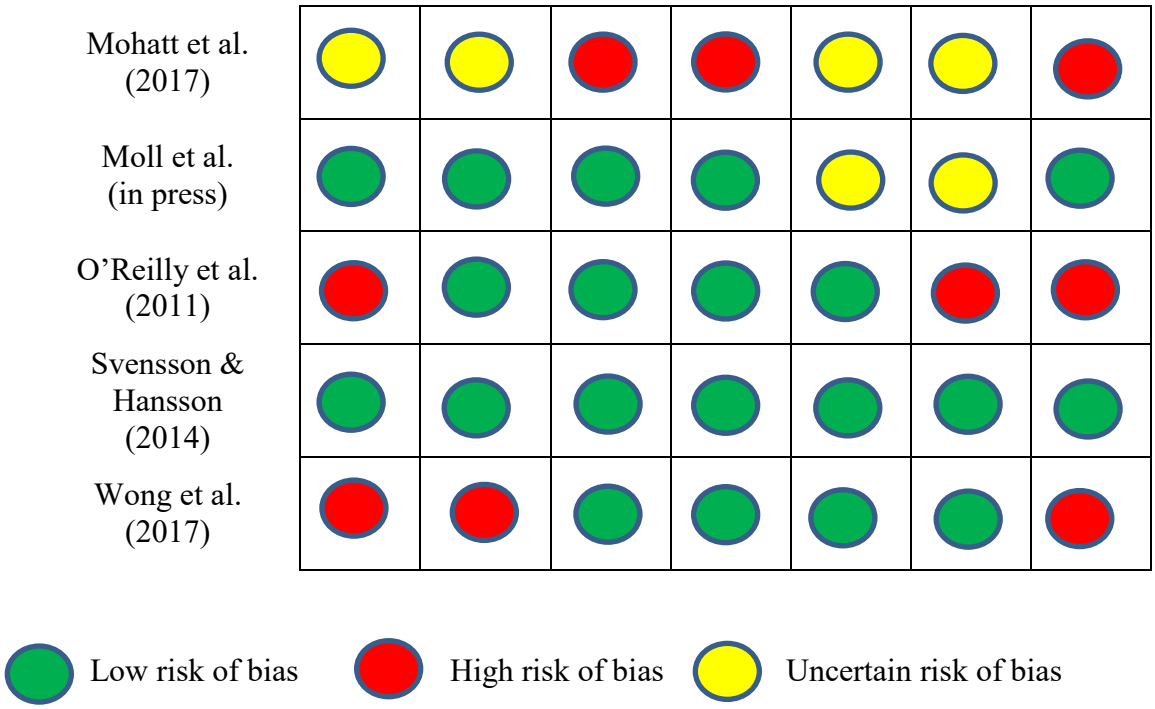


Figure 2. Risk of bias by study and category.

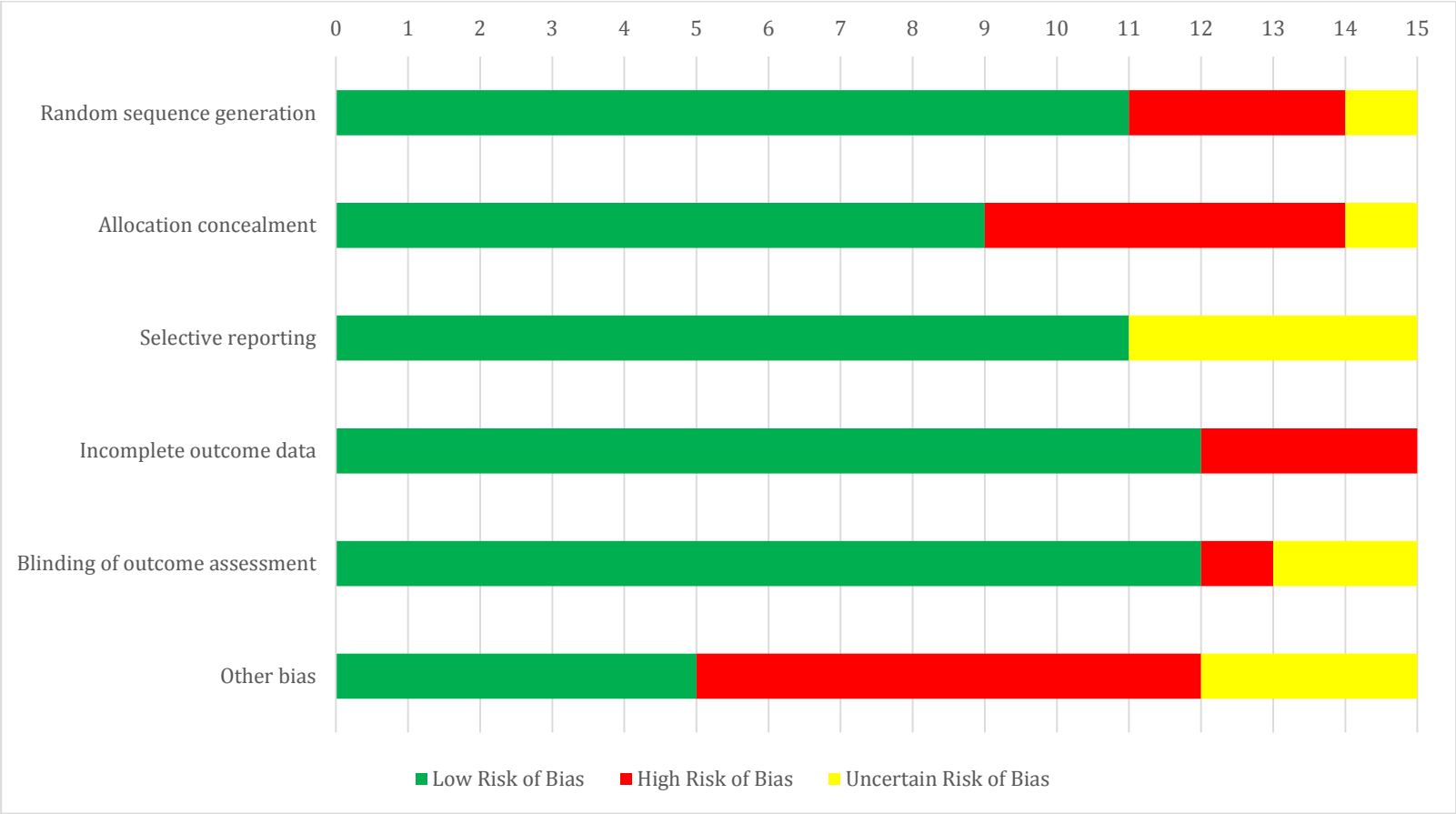


Figure 3. Risk of bias by category for all included studies.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Burns et al. (2017)	0.573	0.240	0.058	0.103	1.043	2.389	0.017
Davies et al. (2018)	1.113	0.372	0.138	0.385	1.842	2.996	0.003
Jensen et al. (2016)	0.219	0.126	0.016	-0.028	0.466	1.741	0.082
Jorm et al. (2004)	0.337	0.088	0.008	0.164	0.510	3.822	0.000
Jorm et al. (2010a)	0.867	0.395	0.156	0.094	1.641	2.198	0.028
Jorm et al. (2010b)	0.455	0.306	0.094	-0.146	1.055	1.484	0.138
Kitchener & Jorm (2004)	0.143	0.178	0.032	-0.206	0.492	0.802	0.422
Lipson et al. (2014)	0.200	0.085	0.007	0.033	0.367	2.346	0.019
Massey et al. (2014)	0.933	0.239	0.057	0.464	1.402	3.898	0.000
Moffitt et al. (2014)	0.339	0.286	0.082	-0.222	0.901	1.185	0.236
Mohatt et al. (2017)	0.585	0.321	0.103	-0.043	1.214	1.825	0.068
Moll et al. (in press)	0.209	0.144	0.021	-0.074	0.492	1.448	0.148
O'Reilly et al. (2011)	0.663	0.436	0.190	-0.191	1.518	1.521	0.128
Svensson & Hansson (2014)	0.237	0.122	0.015	-0.002	0.476	1.939	0.052
Wong et al. (2017)	0.282	0.121	0.015	0.045	0.518	2.332	0.020
	0.335	0.054	0.003	0.230	0.441	6.246	0.000

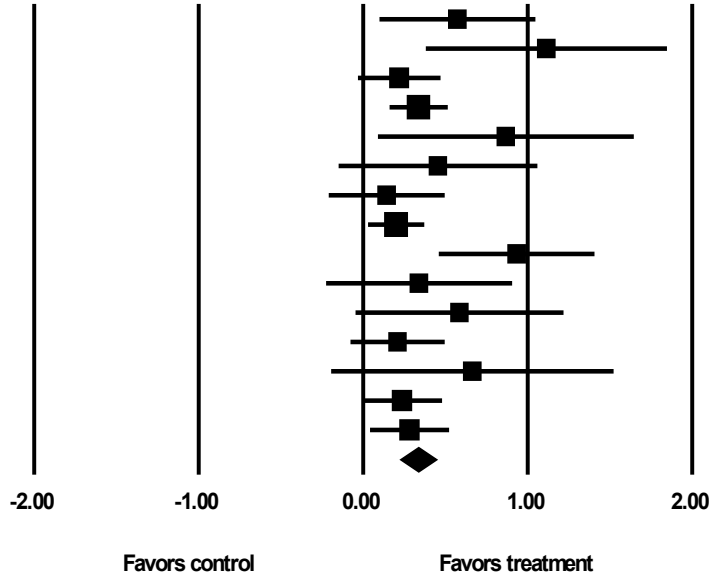


Figure 4. Forest plot of overall effect of trainees' knowledge, attitude, and behavior from baseline to latest follow-up.

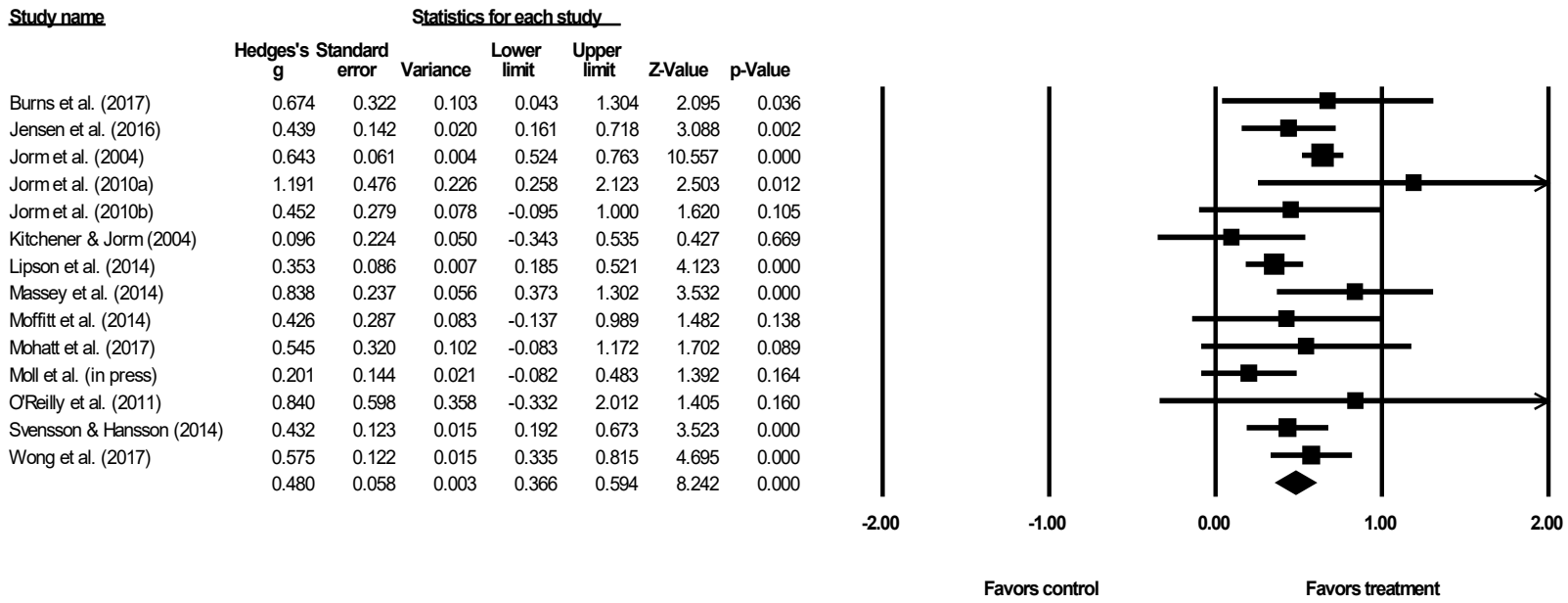


Figure 5. Forest plot of overall effect of trainees' knowledge from baseline to latest follow-up.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Burns et al. (2017)	0.975	0.180	0.032	0.622	1.327	5.418	0.000
Jorm et al. (2010a)	1.111	0.483	0.233	0.165	2.057	2.302	0.021
Jorm et al. (2010b)	0.311	0.270	0.073	-0.218	0.840	1.154	0.249
Moffitt et al. (2014)	0.426	0.287	0.083	-0.137	0.989	1.482	0.138
Moll et al. (in press)	0.230	0.144	0.021	-0.053	0.513	1.594	0.111
Wong et al. (2017)	0.604	0.123	0.015	0.364	0.845	4.929	0.000
	0.561	0.137	0.019	0.294	0.829	4.108	0.000

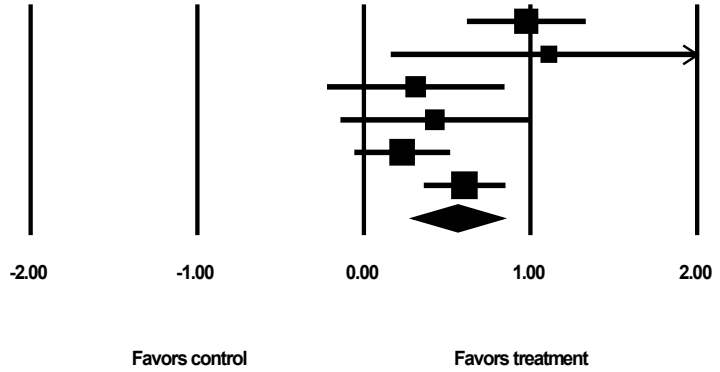


Figure 6. Forest plot of overall effect of trainees' knowledge from baseline to immediate post-intervention.

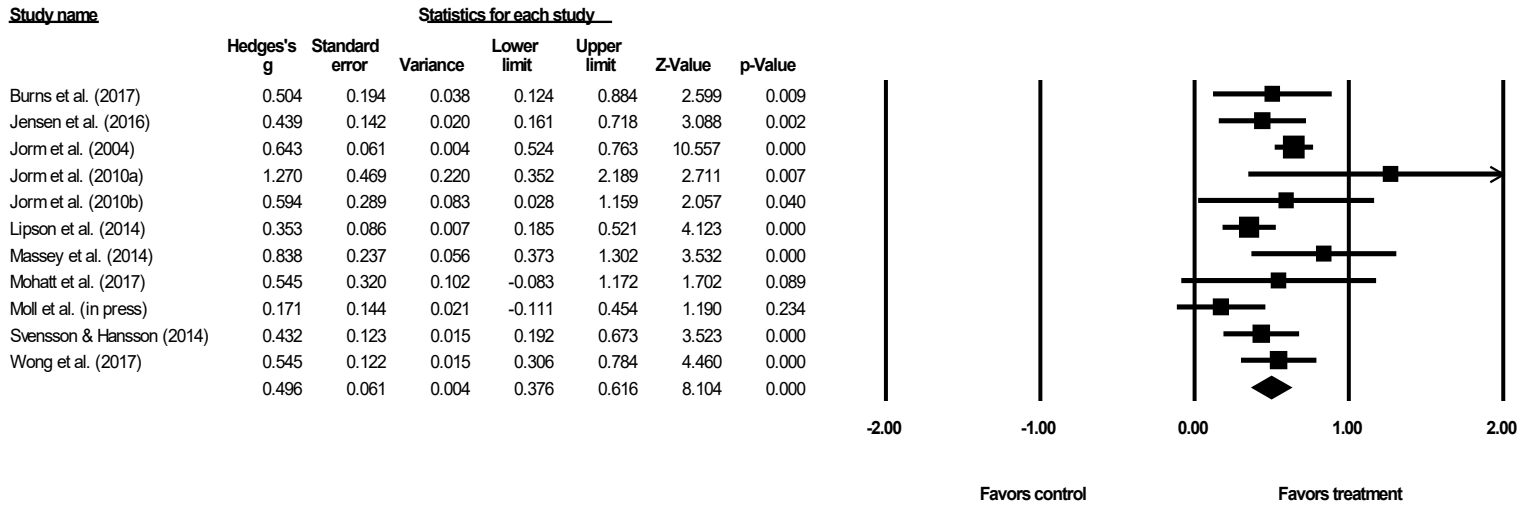


Figure 7. Forest plot of overall effect of trainees' knowledge from baseline to latest follow-up, without immediate post-intervention data.

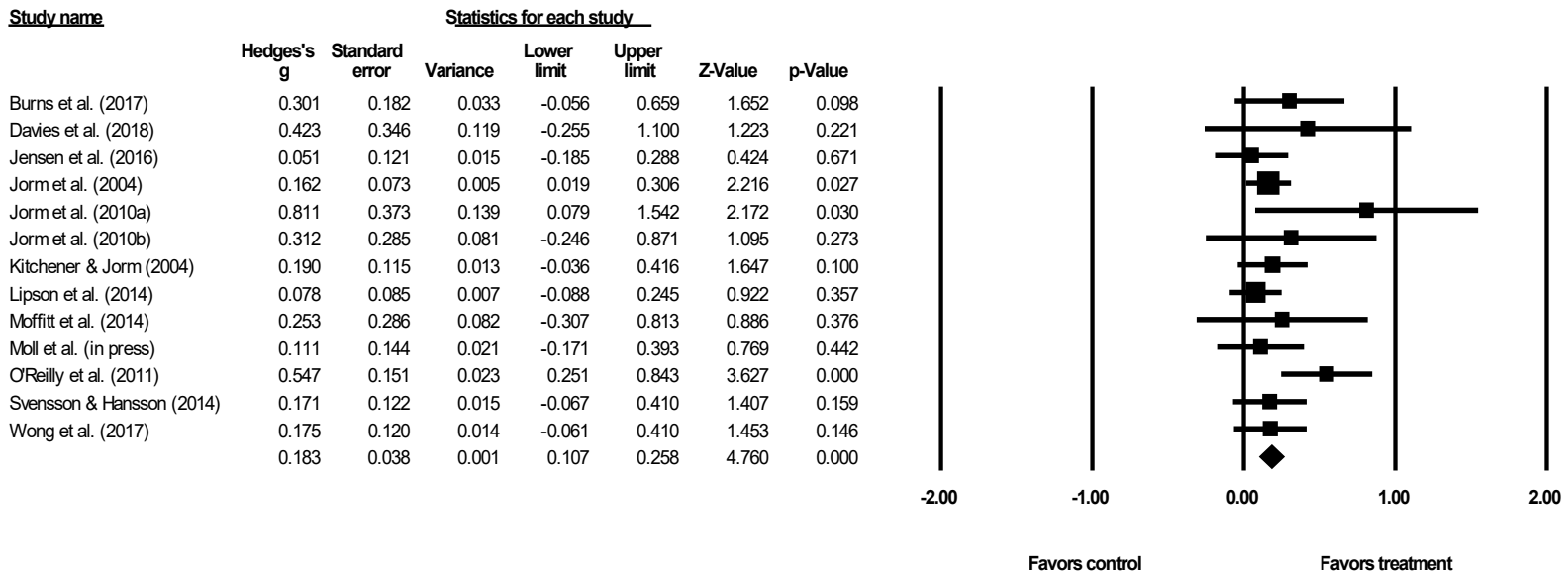


Figure 8. Forest plot of overall effect of trainees' attitude from baseline to latest follow-up.

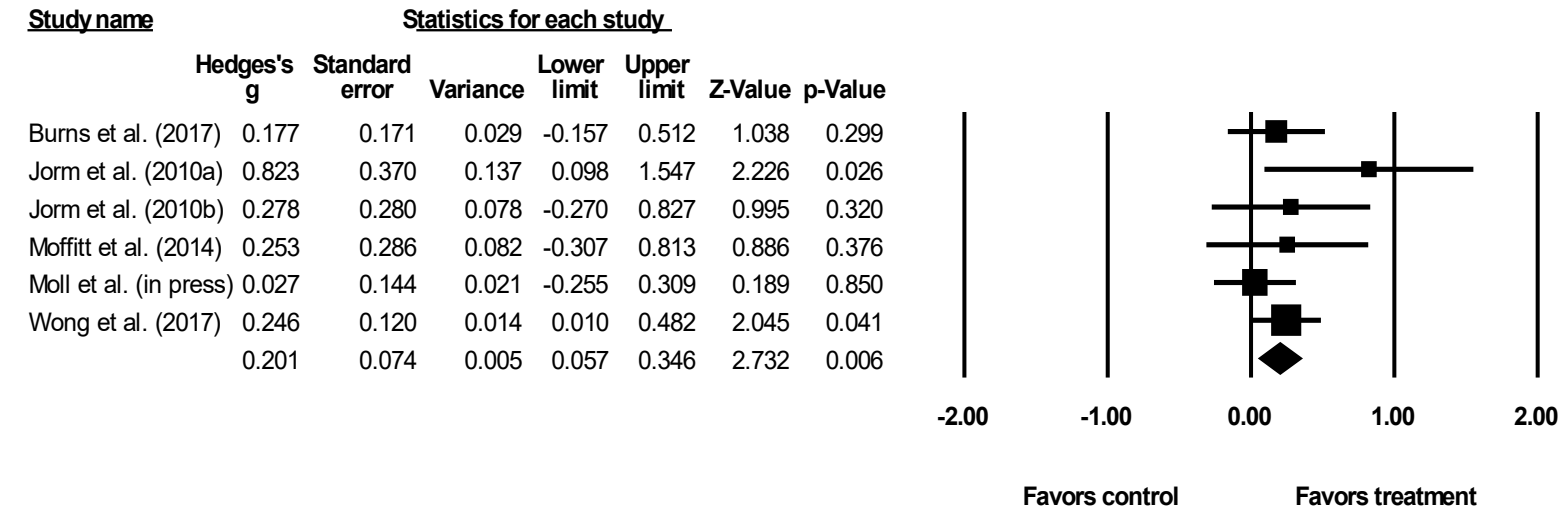


Figure 9. Forest plot of overall effect of trainees' attitude from baseline to immediate post-intervention.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Burns et al. (2017)	0.425	0.193	0.037	0.047	0.804	2.201	0.028
Davies et al. (2018)	0.423	0.346	0.119	-0.255	1.100	1.223	0.221
Jensen et al. (2016)	0.051	0.121	0.015	-0.185	0.288	0.424	0.671
Jorm et al. (2004)	0.162	0.073	0.005	0.019	0.306	2.216	0.027
Jorm et al. (2010a)	0.799	0.377	0.142	0.060	1.537	2.119	0.034
Jorm et al. (2010b)	0.346	0.290	0.084	-0.223	0.915	1.193	0.233
Kitchener & Jorm (2004)	0.190	0.115	0.013	-0.036	0.416	1.647	0.100
Lipson et al. (2014)	0.078	0.085	0.007	-0.088	0.245	0.922	0.357
Moll et al. (in press)	0.194	0.144	0.021	-0.088	0.477	1.348	0.178
O'Reilly et al. (2011)	0.547	0.151	0.023	0.251	0.843	3.627	0.000
Svensson & Hansson (2014)	0.171	0.122	0.015	-0.067	0.410	1.407	0.159
Wong et al. (2017)	0.103	0.120	0.014	-0.132	0.338	0.859	0.391
	0.194	0.044	0.002	0.107	0.280	4.399	0.000

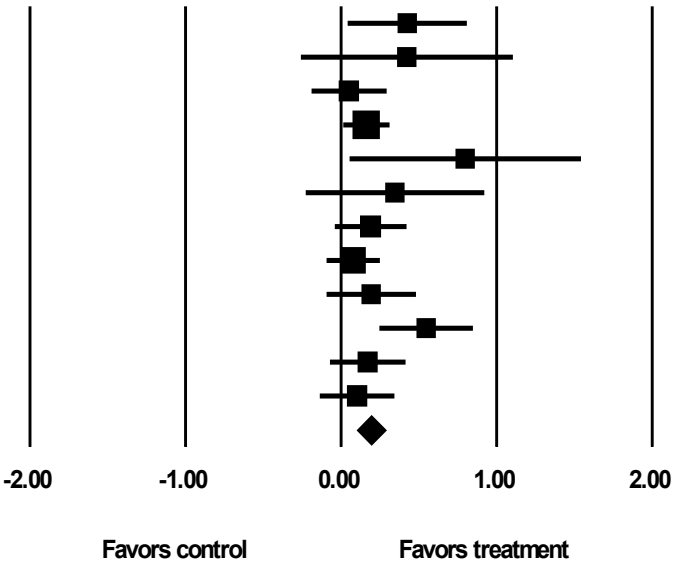


Figure 10. Forest plot of overall effect of trainees' attitude from baseline to latest follow-up, without immediate post-intervention data.

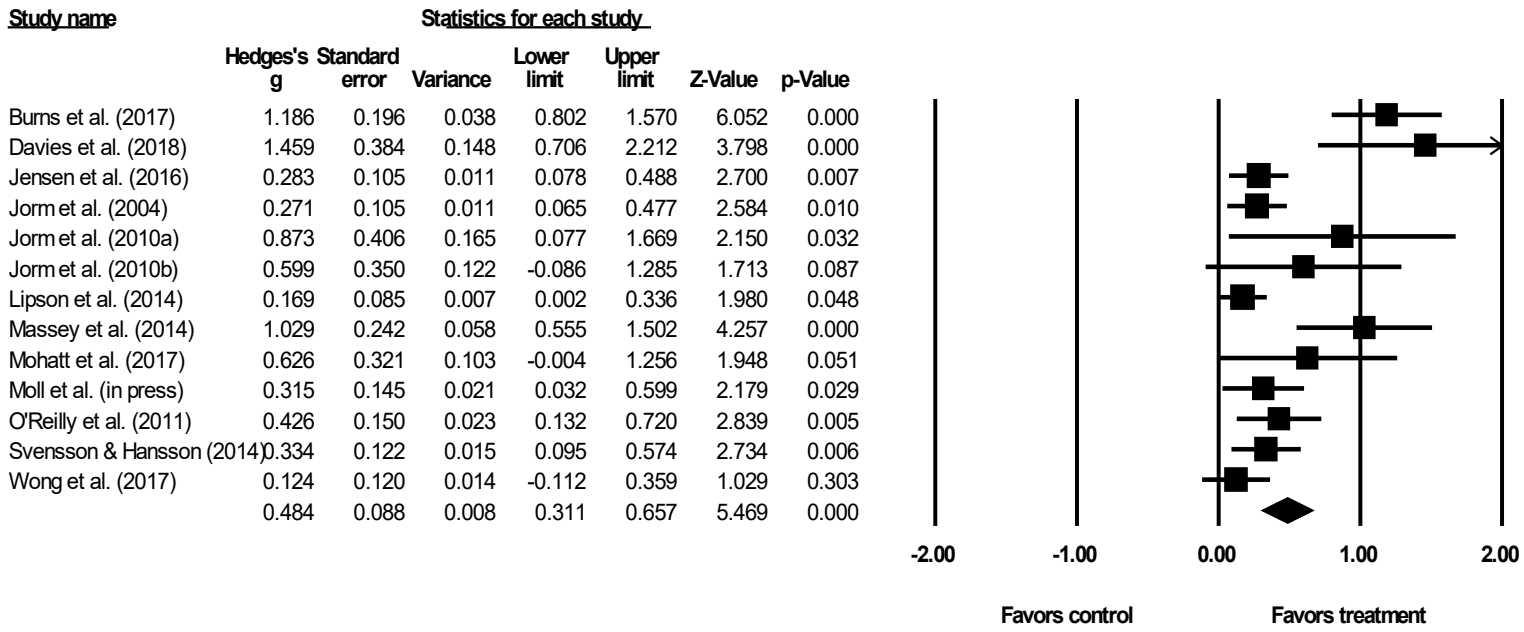


Figure 11. Forest plot of overall effect of trainees' behavior from baseline to latest follow-up.

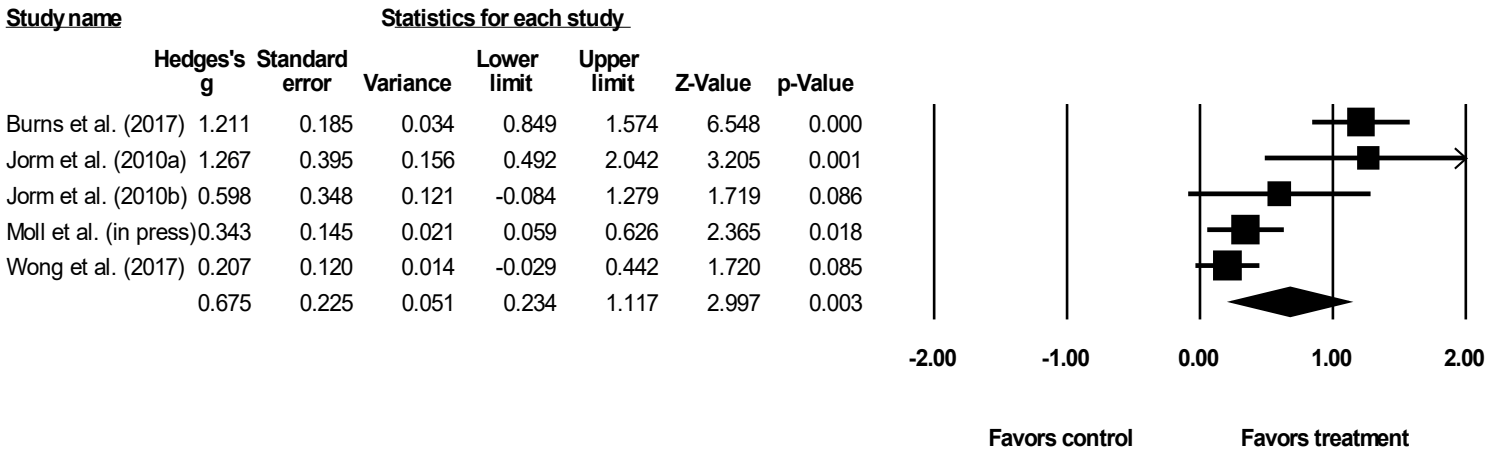


Figure 12. Forest plot of overall effect of trainees' behavior from baseline to immediate post-intervention.

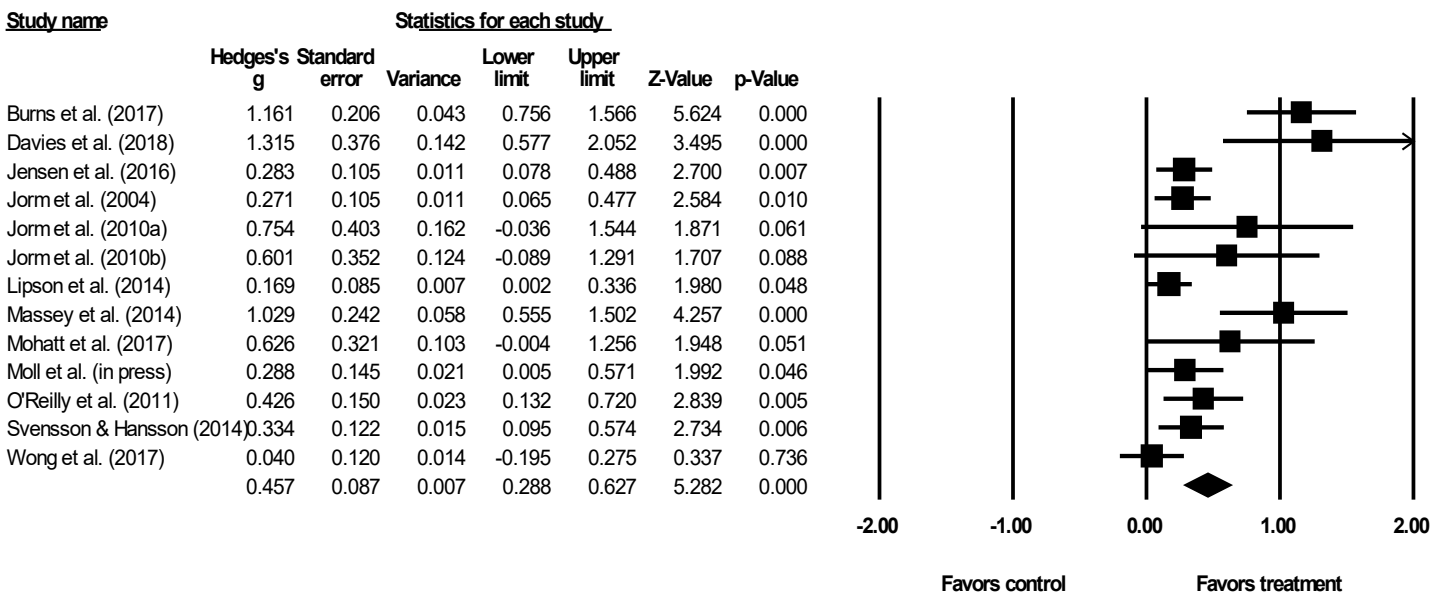


Figure 13. Forest plot of overall effect of trainees' behavior from baseline to latest follow-up, without immediate post-intervention data.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Davies et al. (2018)	0.320	0.344	0.118	-0.354	0.994	0.930	0.353
Jorm et al. (2010a)	0.125	0.376	0.141	-0.612	0.862	0.333	0.739
Jorm et al. (2010b)	0.148	0.302	0.091	-0.443	0.739	0.490	0.624
Kitchener & Jorm (2004)	0.209	0.115	0.013	-0.017	0.435	1.809	0.070
Lipson et al. (2014)	0.048	0.085	0.007	-0.119	0.214	0.561	0.575
	0.115	0.065	0.004	-0.012	0.241	1.775	0.076

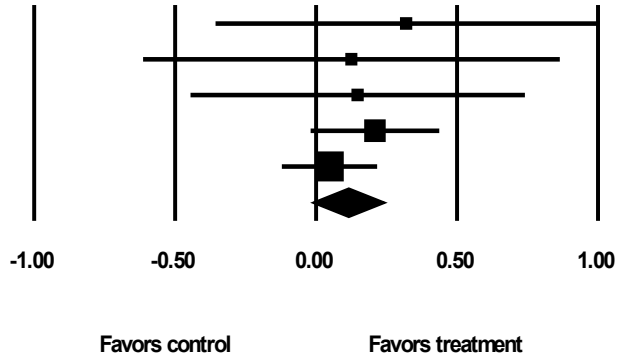


Figure 14. Forest plot of overall effect of trainees' distress from baseline to latest follow-up.

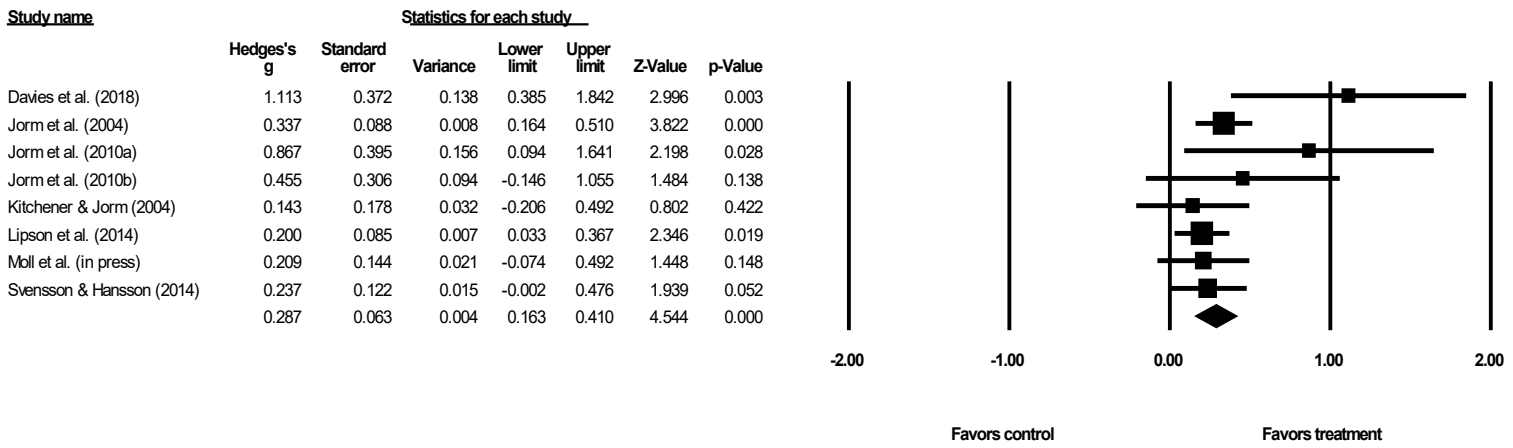


Figure 15. Forest plot of overall effect of trainees' knowledge, attitude, and behavior from baseline to latest follow-up with only low bias studies.

Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Jorm et al. (2004)	0.643	0.061	0.004	0.524	0.763	10.557	0.000
Jorm et al. (2010a)	1.191	0.476	0.226	0.258	2.123	2.503	0.012
Jorm et al. (2010b)	0.452	0.279	0.078	-0.095	1.000	1.620	0.105
Kitchener & Jorm (2004)	0.096	0.224	0.050	-0.343	0.535	0.427	0.669
Lipson et al. (2014)	0.353	0.086	0.007	0.185	0.521	4.123	0.000
Moll et al. (in press)	0.201	0.144	0.021	-0.082	0.483	1.392	0.164
Svensson & Hansson (2014)	0.432	0.123	0.015	0.192	0.673	3.523	0.000
	0.420	0.092	0.009	0.239	0.601	4.546	0.000

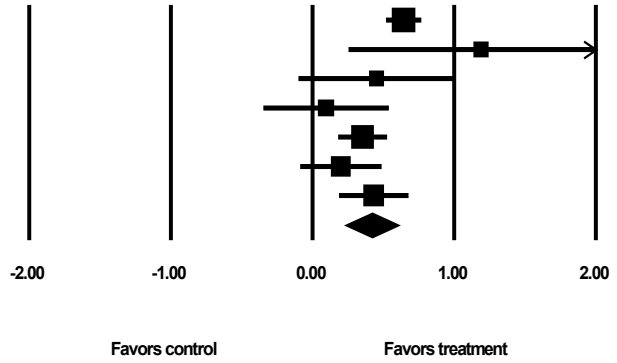


Figure 16. Forest plot of overall effect of trainees' knowledge from baseline to latest follow-up with only low bias studies.

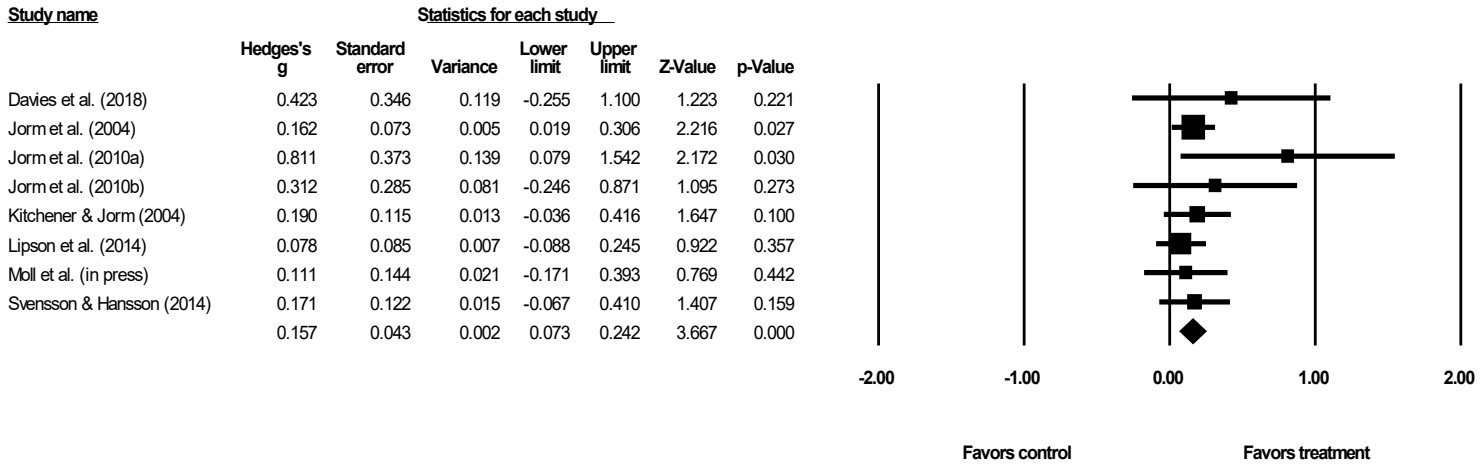


Figure 17. Forest plot of overall effect of trainees' attitude from baseline to latest follow-up with only low bias studies.

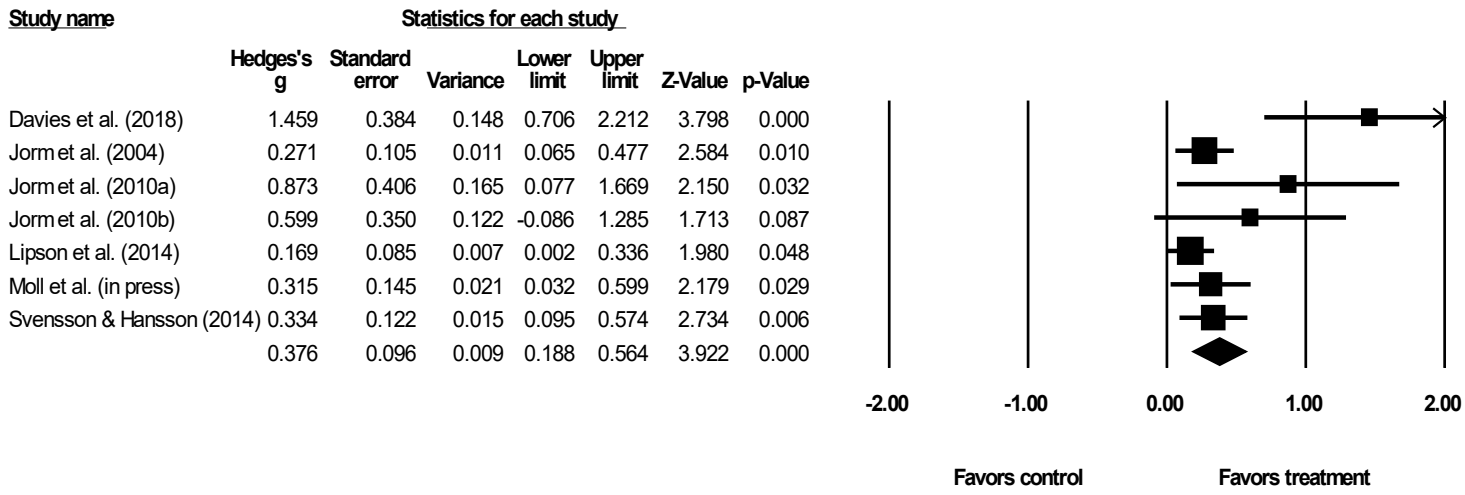


Figure 18. Forest plot of overall effect of trainees' behavior from baseline to latest follow-up with only low bias studies.

Group by Study design	Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
nRCT	Massey et al. (2014)	0.933	0.239	0.057	0.464	1.402	3.898	0.000
nRCT	Moffitt et al. (2014)	0.339	0.286	0.082	-0.222	0.901	1.185	0.236
nRCT	O'Reilly et al. (2011)	0.663	0.436	0.190	-0.191	1.518	1.521	0.128
nRCT	Wong et al. (2017)	0.282	0.121	0.015	0.045	0.518	2.332	0.020
nRCT		0.457	0.121	0.015	0.220	0.695	3.771	0.000
RCT	Burns et al. (2017)	0.573	0.240	0.058	0.103	1.043	2.389	0.017
RCT	Davies et al. (2018)	1.113	0.372	0.138	0.385	1.842	2.996	0.003
RCT	Jensen et al. (2016)	0.219	0.126	0.016	-0.028	0.466	1.741	0.082
RCT	Jorm et al. (2004)	0.337	0.088	0.008	0.164	0.510	3.822	0.000
RCT	Jorm et al. (2010a)	0.867	0.395	0.156	0.094	1.641	2.198	0.028
RCT	Jorm et al. (2010b)	0.455	0.306	0.094	-0.146	1.055	1.484	0.138
RCT	Kitchener & Jorm (2004)	0.143	0.178	0.032	-0.206	0.492	0.802	0.422
RCT	Lipson et al. (2014)	0.200	0.085	0.007	0.033	0.367	2.346	0.019
RCT	Mohatt et al. (2017)	0.585	0.321	0.103	-0.043	1.214	1.825	0.068
RCT	Moll et al. (in press)	0.209	0.144	0.021	-0.074	0.492	1.448	0.148
RCT	Svensson & Hansson (2014)	0.237	0.122	0.015	-0.002	0.476	1.939	0.052
RCT		0.306	0.060	0.004	0.188	0.424	5.088	0.000
Overall		0.345	0.066	0.004	0.215	0.474	5.216	0.000

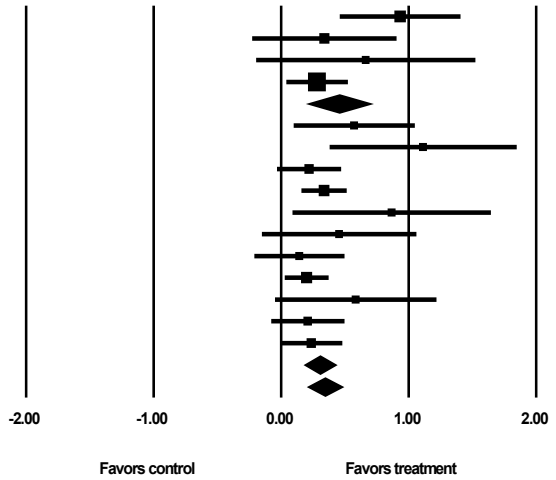
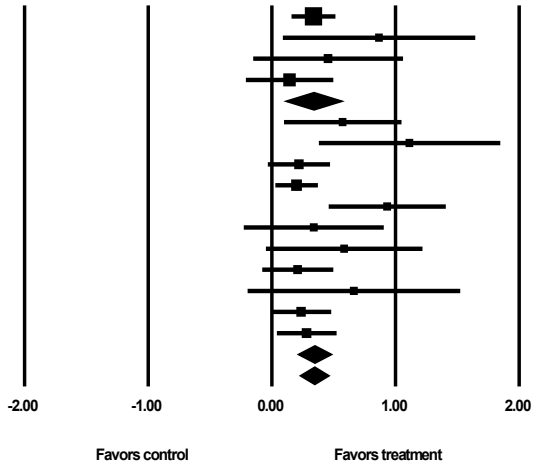


Figure 19. Forest plot of nRCTs vs. RCTs moderation effect.

Group by Investigator allegiance proxy	Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
JK	Jorm et al. (2004)	0.337	0.088	0.008	0.164	0.510	3.822	0.000
JK	Jorm et al. (2010a)	0.867	0.395	0.156	0.094	1.641	2.198	0.028
JK	Jorm et al. (2010b)	0.455	0.306	0.094	-0.146	1.055	1.484	0.138
JK	Kitchener & Jorm (2004)	0.143	0.178	0.032	-0.206	0.492	0.802	0.422
JK		0.338	0.113	0.013	0.117	0.560	2.991	0.003
No	Burns et al. (2017)	0.573	0.240	0.058	0.103	1.043	2.389	0.017
No	Davies et al. (2018)	1.113	0.372	0.138	0.385	1.842	2.996	0.003
No	Jensen et al. (2016)	0.219	0.126	0.016	-0.028	0.466	1.741	0.082
No	Lipson et al. (2014)	0.200	0.085	0.007	0.033	0.367	2.346	0.019
No	Massey et al. (2014)	0.933	0.239	0.057	0.464	1.402	3.898	0.000
No	Moffitt et al. (2014)	0.339	0.286	0.082	-0.222	0.901	1.185	0.236
No	Mohatt et al. (2017)	0.585	0.321	0.103	-0.043	1.214	1.825	0.068
No	Moll et al. (in press)	0.209	0.144	0.021	-0.074	0.492	1.448	0.148
No	O'Reilly et al. (2011)	0.663	0.436	0.190	-0.191	1.518	1.521	0.128
No	Svensson & Hansson (2014)	0.237	0.122	0.015	-0.002	0.476	1.939	0.052
No	Wong et al. (2017)	0.282	0.121	0.015	0.045	0.518	2.332	0.020
No		0.345	0.067	0.004	0.215	0.476	5.178	0.000
Overall		0.344	0.057	0.003	0.231	0.456	5.980	0.000



Note. JK: involved Jorm and Kitchener, No: did not involve either founder.

Figure 20. Forest plot of investigator allegiance proxy moderation effect.

Group by Country of research	Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Australia	Burns et al. (2017)	0.573	0.240	0.058	0.103	1.043	2.389	0.017
Australia	Jorm et al. (2004)	0.337	0.088	0.008	0.164	0.510	3.822	0.000
Australia	Jorm et al. (2010a)	0.867	0.395	0.156	0.094	1.641	2.198	0.028
Australia	Jorm et al. (2010b)	0.455	0.306	0.094	-0.146	1.055	1.484	0.138
Australia	Kitchener & Jorm (2004)	0.143	0.178	0.032	-0.206	0.492	0.802	0.422
Australia	O'Reilly et al. (2011)	0.663	0.436	0.190	-0.191	1.518	1.521	0.128
Australia		0.385	0.099	0.010	0.190	0.579	3.879	0.000
Other	Davies et al. (2018)	1.113	0.372	0.138	0.385	1.842	2.996	0.003
Other	Jensen et al. (2016)	0.219	0.126	0.016	-0.028	0.466	1.741	0.082
Other	Lipson et al. (2014)	0.200	0.085	0.007	0.033	0.367	2.346	0.019
Other	Massey et al. (2014)	0.933	0.239	0.057	0.464	1.402	3.898	0.000
Other	Moffitt et al. (2014)	0.339	0.286	0.082	-0.222	0.901	1.185	0.236
Other	Mohatt et al. (2017)	0.585	0.321	0.103	-0.043	1.214	1.825	0.068
Other	Moll et al. (in press)	0.209	0.144	0.021	-0.074	0.492	1.448	0.148
Other	Svensson & Hansson (2014)	0.237	0.122	0.015	-0.002	0.476	1.939	0.052
Other	Wong et al. (2017)	0.282	0.121	0.015	0.045	0.518	2.332	0.020
Other		0.320	0.068	0.005	0.187	0.452	4.715	0.000
Overall		0.340	0.056	0.003	0.231	0.450	6.081	0.000

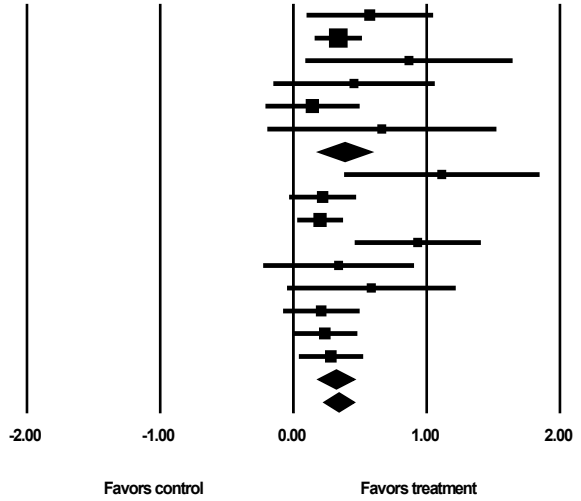


Figure 21. Forest plot of country of research moderation effect.

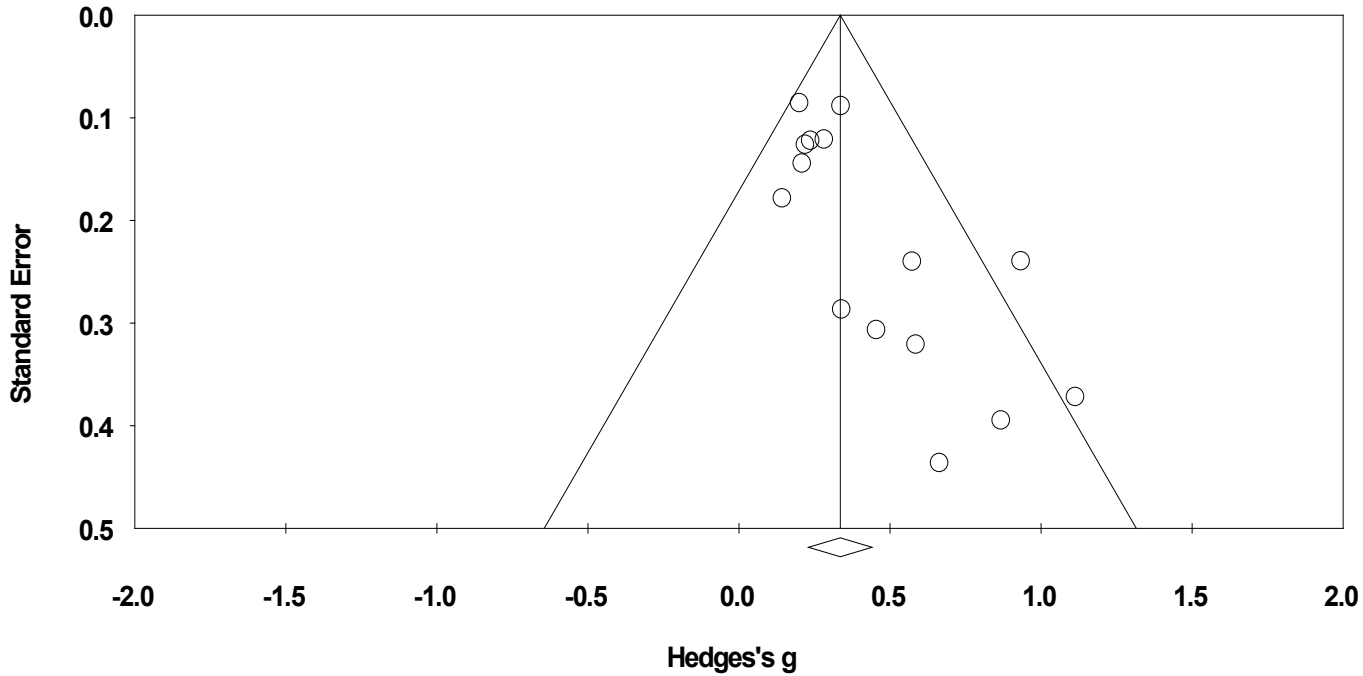


Figure 22. Funnel plot for overall effect of trainees' knowledge, attitude, and behavior.

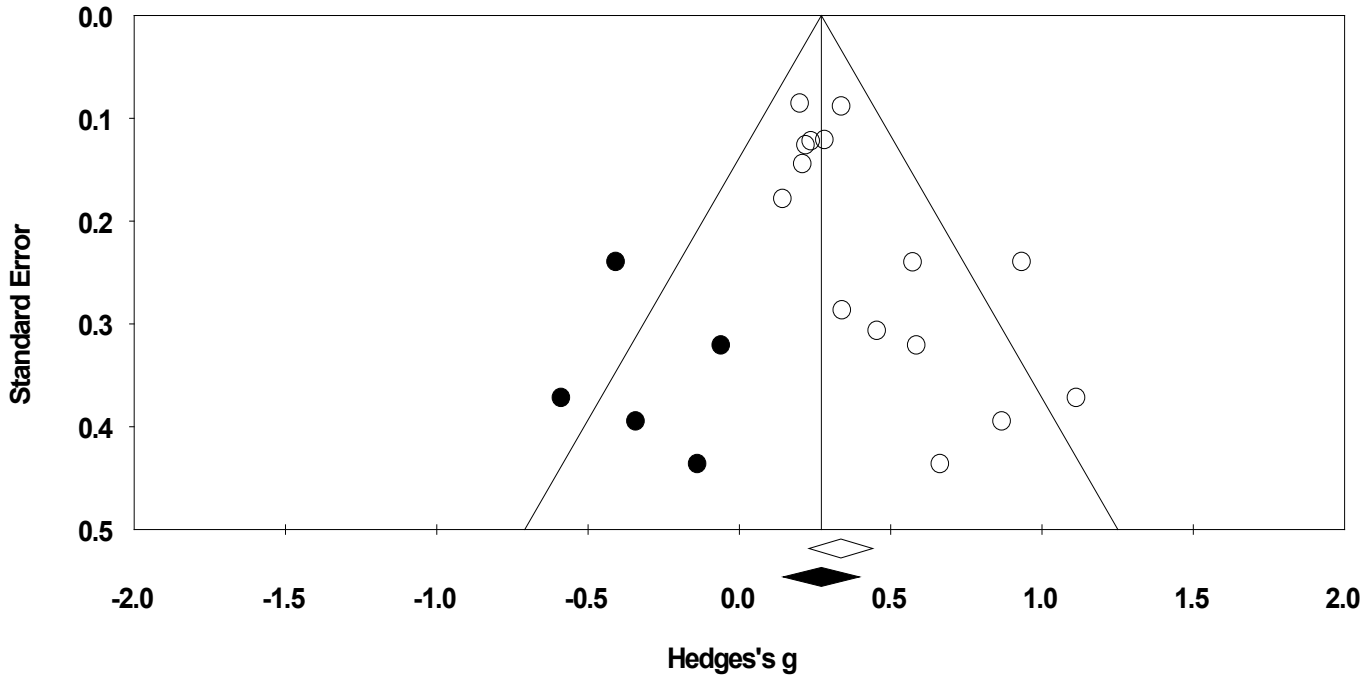


Figure 23. Funnel plot for overall effect of trainees' knowledge, attitude, and behavior with missing studies.

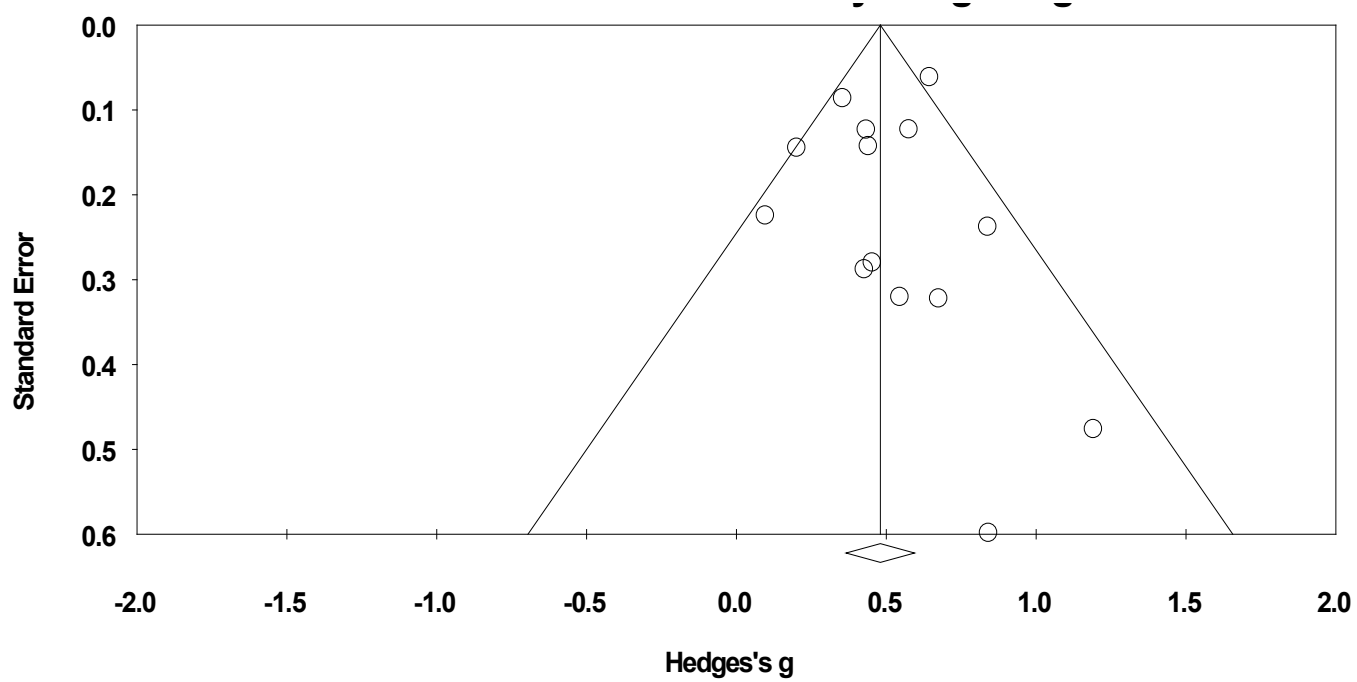


Figure 24. Funnel plot for trainees' knowledge.

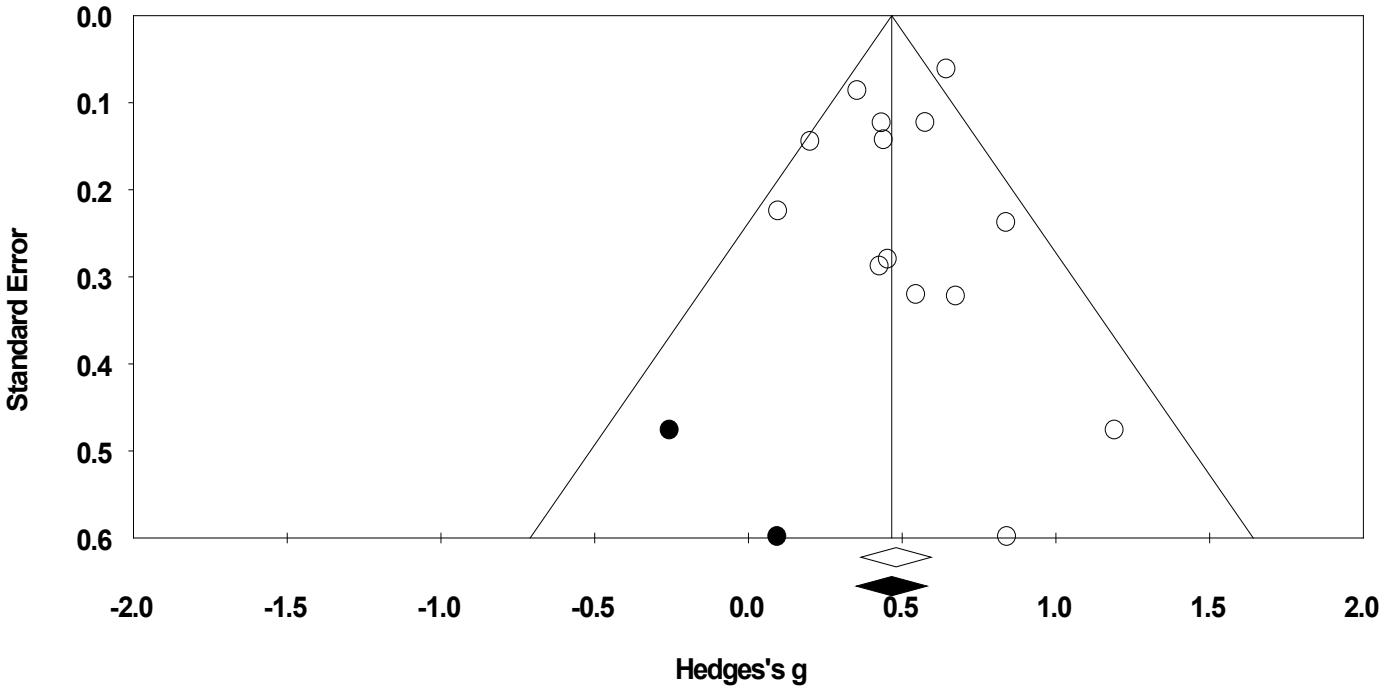


Figure 25. Funnel plot for trainees' knowledge with missing studies.

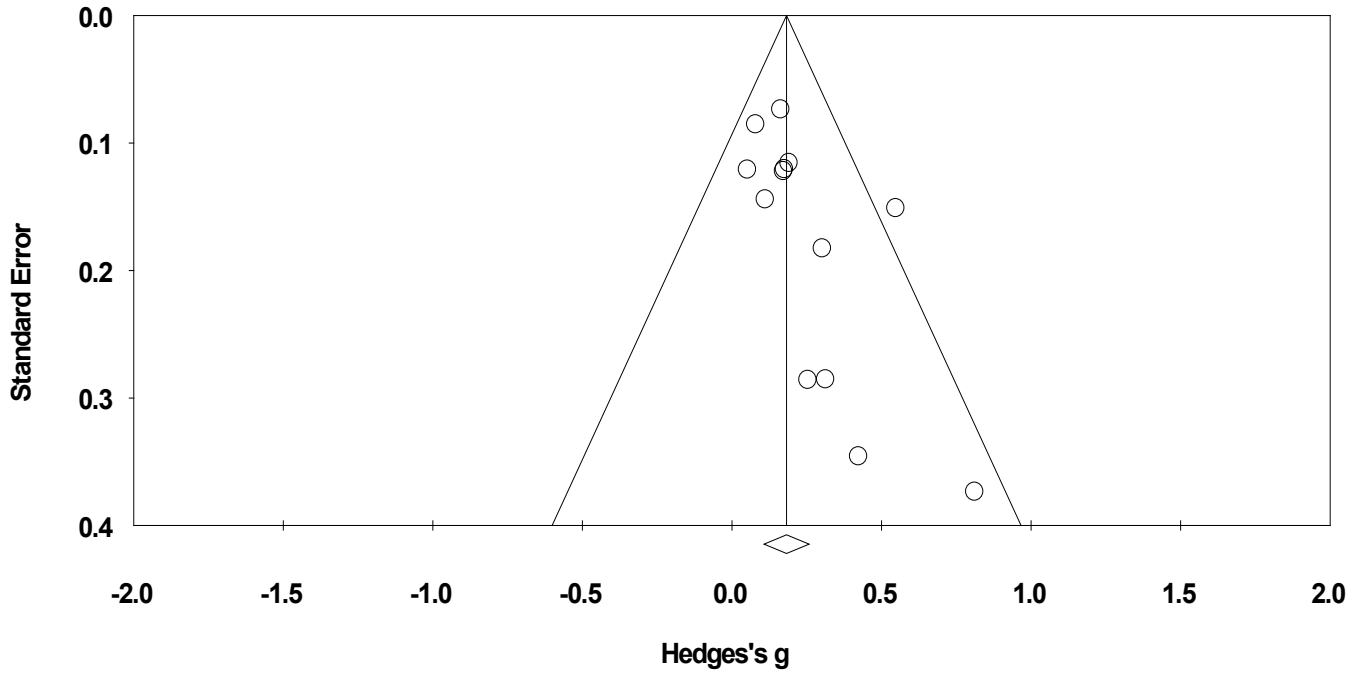


Figure 26. Funnel plot for trainees' attitude.

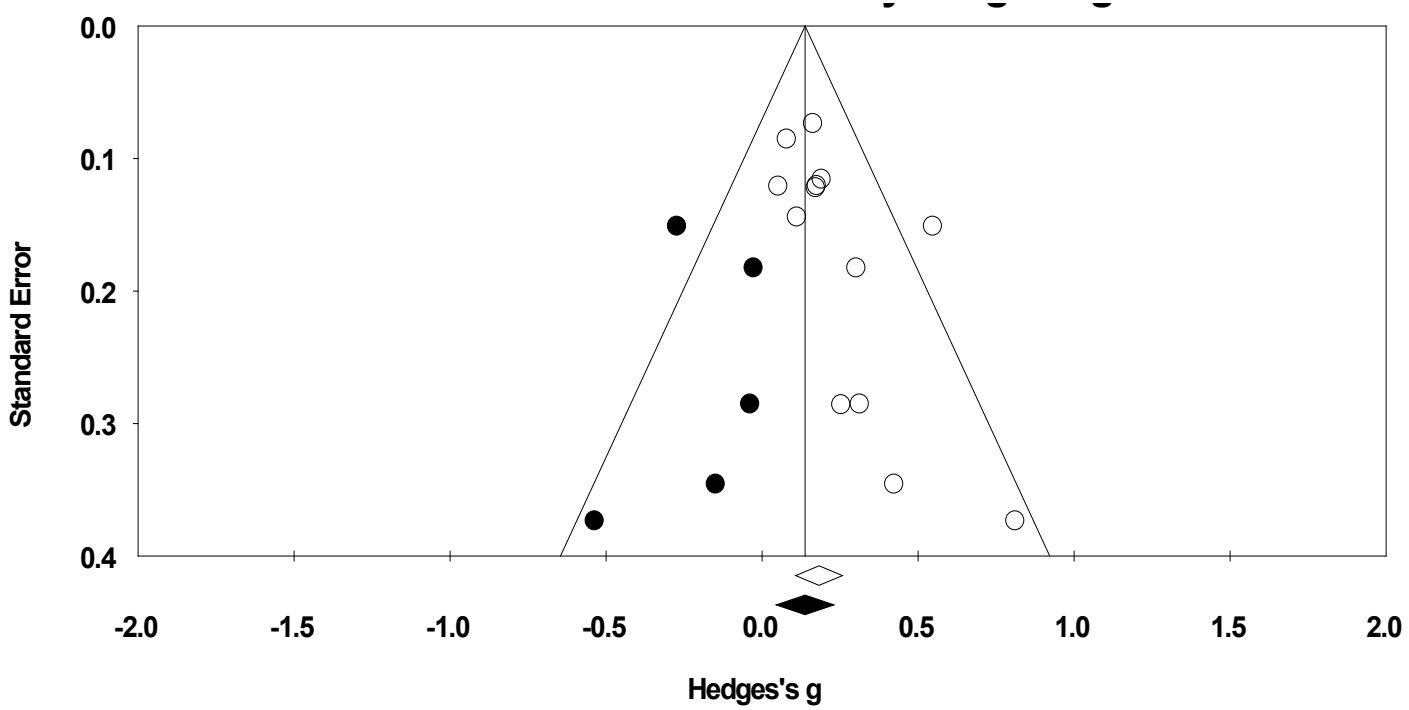


Figure 27. Funnel plot for trainees' attitude with missing studies.

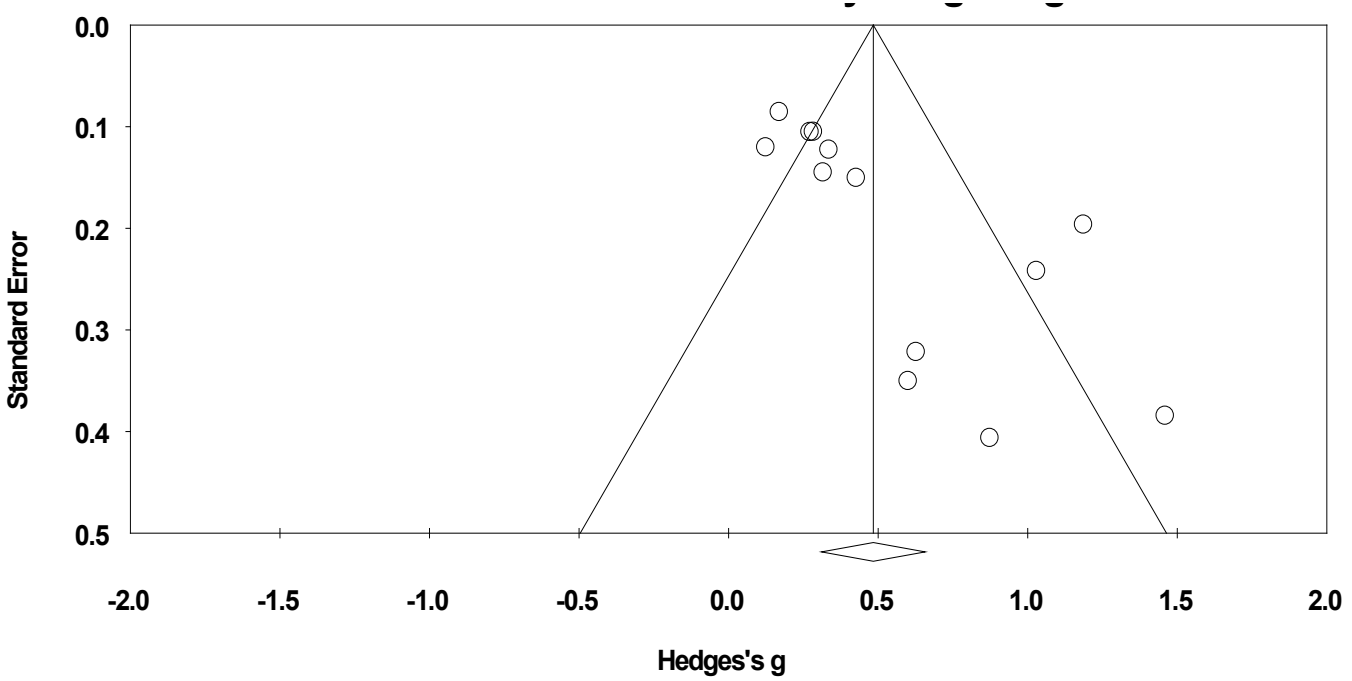


Figure 28. Funnel plot for trainees' behavior.

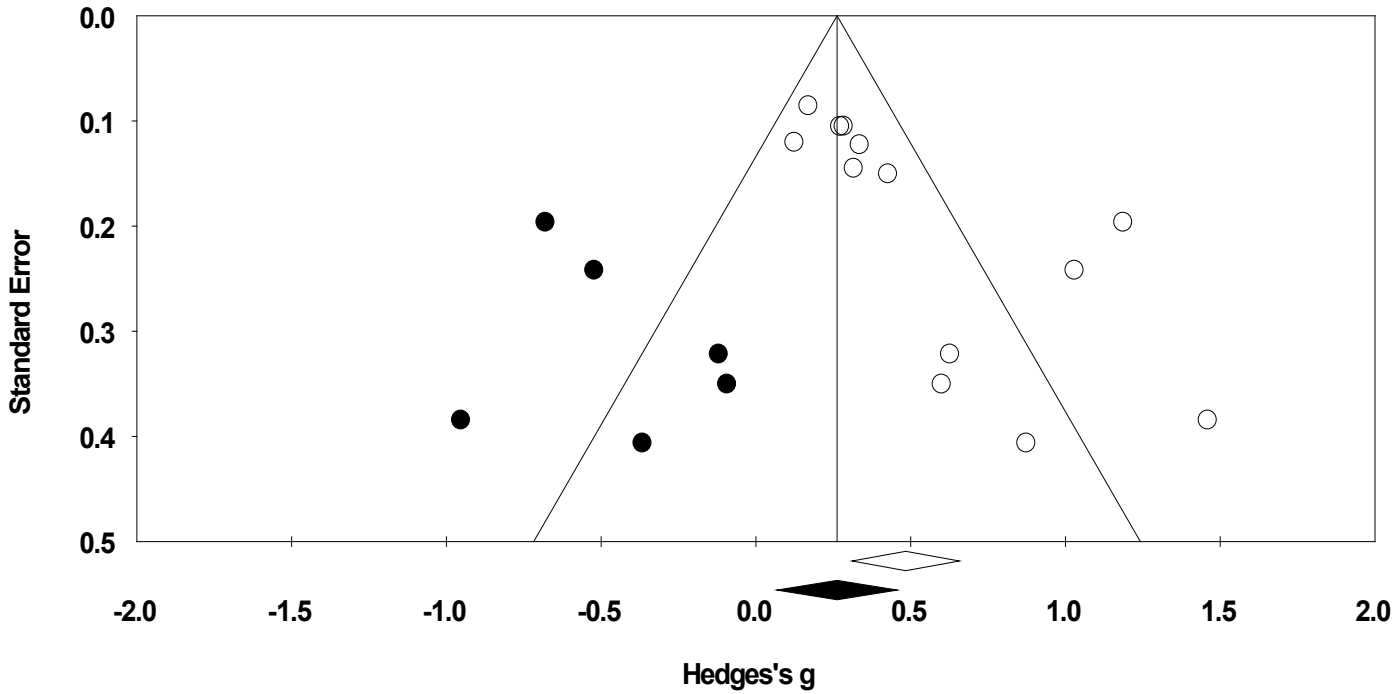


Figure 29. Funnel plot for trainees' behavior with missing studies.

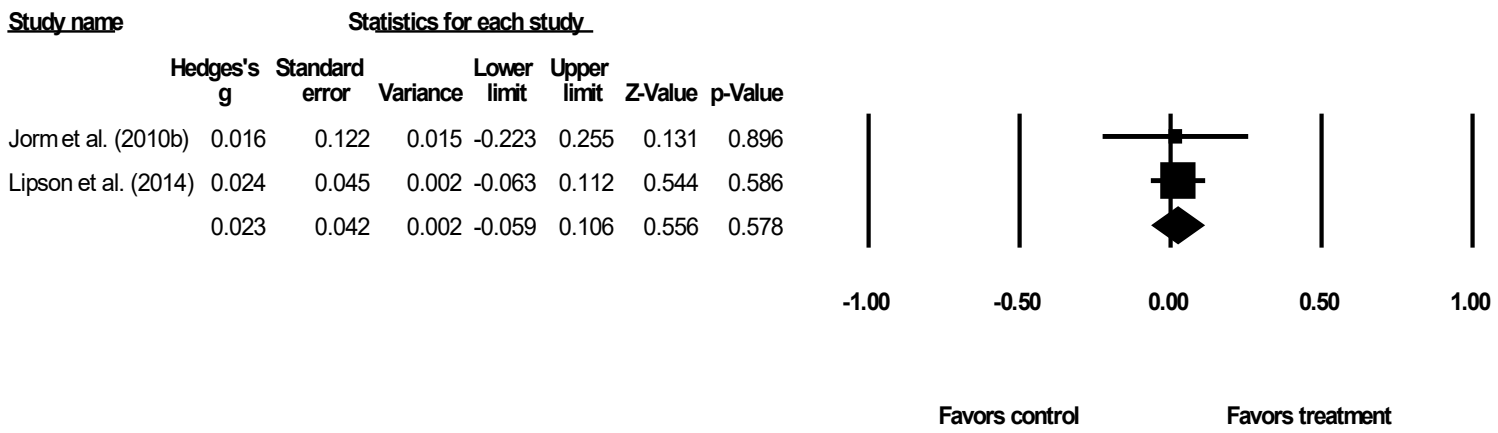


Figure 30. Forest plot of overall effect of recipients' knowledge from baseline to latest follow-up.

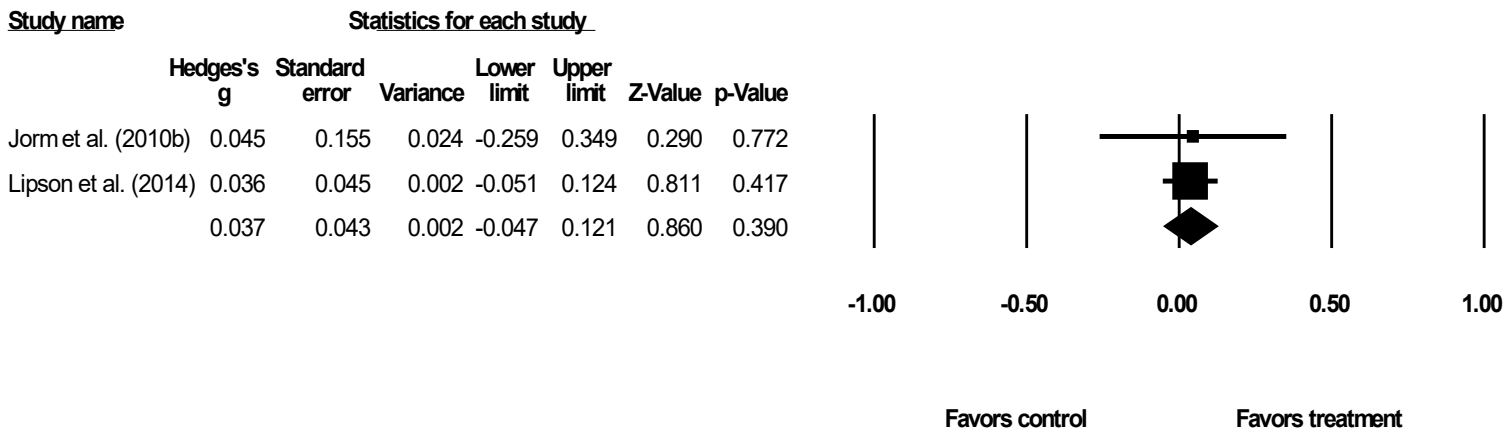


Figure 31. Forest plot of overall effect of recipients' attitude from baseline to latest follow-up.

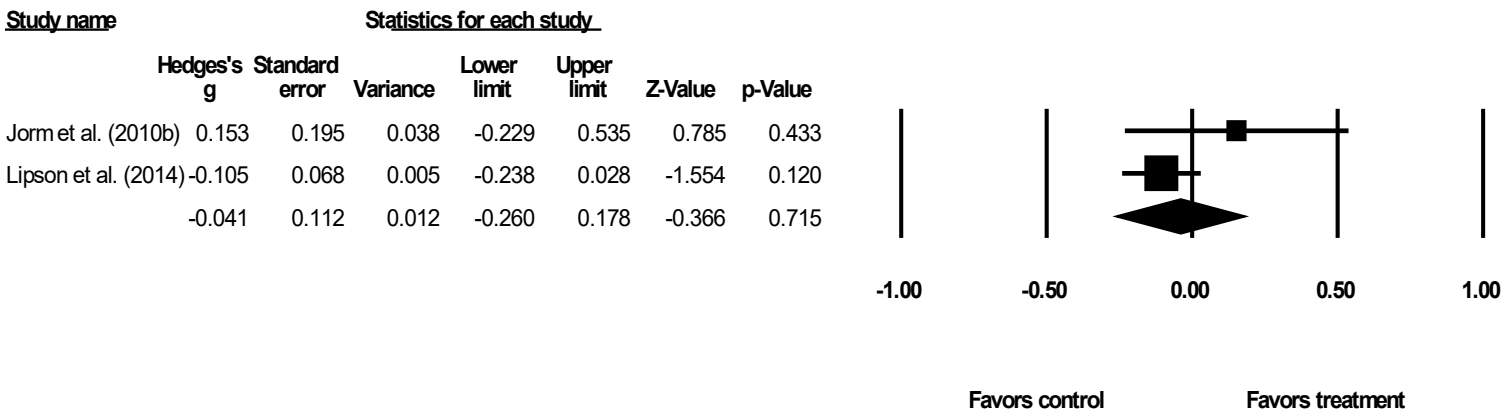


Figure 32. Forest plot of overall effect of recipients' help received or sought from baseline to latest follow-up.

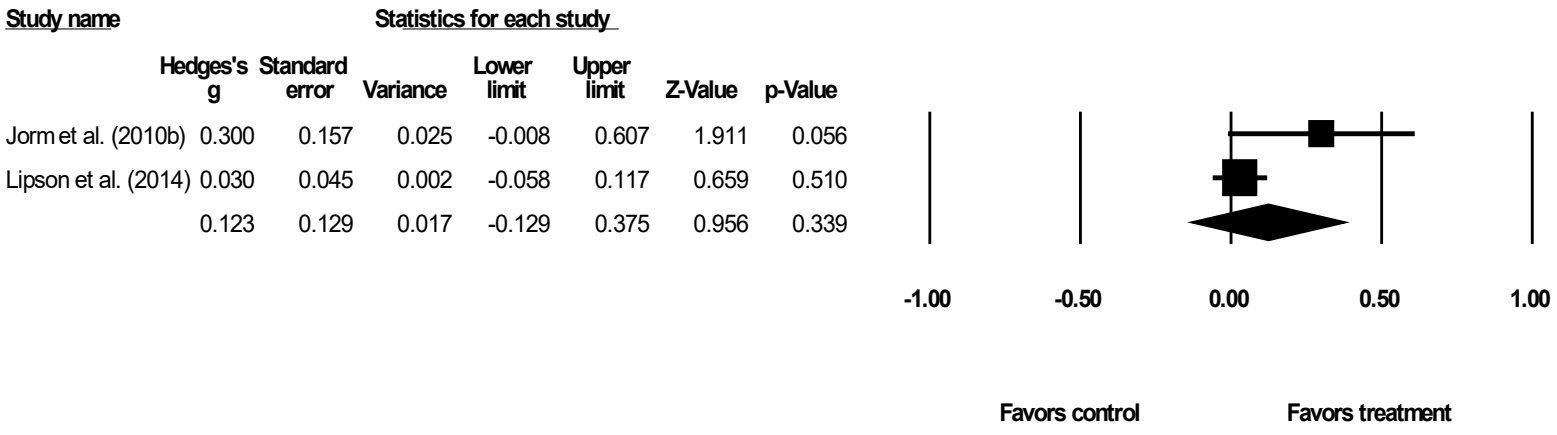


Figure 33. Forest plot of overall effect of recipients' distress from baseline to latest follow-up.

Appendix A: PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

Appendix B: Search Syntax & Journal Search

Names of Databases: PsycINFO, PsycARTICLES, Psychology Database, Consumer Health Complete, PubMed, and Scopus + Dissertations & Theses + journal[a] table of contents sections

Dates: May 2017 - March 2018

Initials of person who ran search: AM

1. MHFA OR mental health first aid OR mental health gatekeeper AND training OR mental health training OR mental health education OR mental health literacy

Delimited to: peer-reviewed journal articles, theses, and dissertations, published in English

Total hits: 12,295 from initial databases + 19 from dissertations and theses databases + 7,328 from table of contents + 33 from MHFA-USA website + 2 (late August search) + 3 (February search) + 3 (March search) = 19,683

[a] *BMC Psychiatry*, *International Journal of Mental Health*, *American Psychologist*, *British Medical Journal*, *Mental Health Practice*, *Psychiatric Services*, *Psychological Services*, *Academic Psychiatry*, *Canadian Journal of Psychiatry*, *Journal of Social Psychiatry*, *Australian Journal of Rural Health*, *Advances in Mental Health*, *Journal of Public Mental Health*, *International Journal of Mental Health Systems*, *Australasian Psychiatry*, *International Review of Psychiatry*, *International Journal of Mental Health Nursing*, *Journal of Mental Health Training, Education and Practice*

Appendix C: Cochrane Risk of Bias Tool

RANDOM SEQUENCE GENERATION **Selection bias (biased allocation to interventions) due to inadequate generation of a randomised sequence.**	
Criteria for a judgement of 'Low risk' of bias.	The investigators describe a random component in the sequence generation process such as: • Referring to a random number table; • Using a computer random number generator; • Coin tossing; • Shuffling cards or envelopes; • Throwing dice; • Drawing of lots; • Minimization*. *Minimization may be implemented without a random element, and this is considered to be equivalent to being random.
Criteria for the judgement of 'High risk' of bias.	The investigators describe a non-random component in the sequence generation process. Usually, the description would involve some systematic, non-random approach, for example: • Sequence generated by odd or even date of birth; • Sequence generated by some rule based on date (or day) of admission; • Sequence generated by some rule based on hospital or clinic record number. Other non-random approaches happen much less frequently than the systematic approaches mentioned above and tend to be obvious. They usually involve judgement or some method of non-random categorization of participants, for example: • Allocation by judgement of the clinician; • Allocation by preference of the participant; • Allocation based on the results of a laboratory test or a series of tests; • Allocation by availability of the intervention.
Criteria for the judgement of 'Unclear risk' of bias.	Insufficient information about the sequence generation process to permit judgement of 'Low risk' or 'High risk'.
ALLOCATION CONCEALMENT **Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment.**	
Criteria for a judgement of 'Low risk' of bias.	Participants and investigators enrolling participants could not foresee assignment because one of the following, or an equivalent method, was used to conceal allocation: • Central allocation (including telephone, web-based and pharmacy-controlled randomization);

	• Sequentially numbered drug containers of identical appearance; • Sequentially numbered, opaque, sealed envelopes.
Criteria for the judgement of 'High risk' of bias.	Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias, such as allocation based on: • Using an open random allocation schedule (e.g. a list of random numbers); • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or nonopaque or not sequentially numbered); • Alternation or rotation; • Date of birth; • Case record number; • Any other explicitly unconcealed procedure.
Criteria for the judgement of 'Unclear risk' of bias.	Insufficient information to permit judgement of 'Low risk' or 'High risk'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgement – for example if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.
BLINDING OF PARTICIPANTS AND PERSONNEL **Performance bias due to knowledge of the allocated interventions by participants and personnel during the study.**	
Criteria for a judgement of 'Low risk' of bias.	Any one of the following: • No blinding or incomplete blinding, but the review authors judge that the outcome is not likely to be influenced by lack of blinding; • Blinding of participants and key study personnel ensured, and unlikely that the blinding could have been broken.
Criteria for the judgement of 'High risk' of bias.	Any one of the following: • No blinding or incomplete blinding, and the outcome is likely to be influenced by lack of blinding; • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken, and the outcome is likely to be influenced by lack of blinding.
Criteria for the judgement of 'Unclear risk' of bias.	Any one of the following: • Insufficient information to permit judgement of 'Low risk' or 'High risk'; • The study did not address this outcome.
BLINDING OF OUTCOME ASSESSMENT **Detection bias due to knowledge of the allocated interventions by outcome assessors.**	
Criteria for a judgement of 'Low risk' of bias.	Any one of the following: • No blinding of outcome assessment, but the review authors judge that the outcome measurement is not likely to be influenced by lack of blinding; • Blinding of outcome assessment ensured, and unlikely that the blinding could have been broken.

Criteria for the judgement of 'High risk' of bias.	Any one of the following: • No blinding of outcome assessment, and the outcome measurement is likely to be influenced by lack of blinding; • Blinding of outcome assessment, but likely that the blinding could have been broken, and the outcome measurement is likely to be influenced by lack of blinding.
Criteria for the judgement of 'Unclear risk' of bias.	Any one of the following: • Insufficient information to permit judgement of 'Low risk' or 'High risk'; • The study did not address this outcome.
INCOMPLETE OUTCOME DATA **Attrition bias due to amount, nature or handling of incomplete outcome data.**	
Criteria for a judgement of 'Low risk' of bias.	Any one of the following: • No missing outcome data; • Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to be introducing bias); • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have a clinically relevant impact on the intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have a clinically relevant impact on observed effect size; • Missing data have been imputed using appropriate methods.
Criteria for the judgement of 'High risk' of bias.	Any one of the following: • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size; • 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization; • Potentially inappropriate application of simple imputation.
Criteria for the judgement of 'Unclear risk' of bias.	Any one of the following: • Insufficient reporting of attrition/exclusions to permit judgement of 'Low risk' or 'High risk' (e.g. number randomized not stated, no reasons for missing data provided); • The study did not address this outcome.
SELECTIVE REPORTING	

Reporting bias due to selective outcome reporting.	
Criteria for a judgement of 'Low risk' of bias.	Any of the following: • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way; • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon).
Criteria for the judgement of 'High risk' of bias.	Any one of the following: • Not all of the study's pre-specified primary outcomes have been reported; • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified; • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis; • The study report fails to include results for a key outcome that would be expected to have been reported for such a study.
Criteria for the judgement of 'Unclear risk' of bias.	Insufficient information to permit judgement of 'Low risk' or 'High risk'. It is likely that the majority of studies will fall into this category.
OTHER BIAS **Bias due to problems not covered elsewhere in the table.**	
Criteria for a judgement of 'Low risk' of bias.	The study appears to be free of other sources of bias.
Criteria for the judgement of 'High risk' of bias.	There is at least one important risk of bias. For example, the study: • Had a potential source of bias related to the specific study design used; or • Has been claimed to have been fraudulent; or • Had some other problem.
Criteria for the judgement of 'Unclear risk' of bias.	There may be a risk of bias, but there is either: • Insufficient information to assess whether an important risk of bias exists; or • Insufficient rationale or evidence that an identified problem will introduce bias.